Probabilité *a posteriori*: amélioration d'une mesure de confiance en reconnaissance de la parole

Julie Mauclair, Yannick Estève, Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Le Mans, France
{mauclair,esteve,deleglise}@lium.univ-lemans.fr

ABSTRACT

This paper adresses the word posterior probability used as a confidence measure on speech recognition system. We present a new confidence measure based on the behavior of language model back-off used during the recognition processing. Merging this new confidence measure with word posterior probability allows to obtain a fusion confidence measure, called WP/LMBB, which outperfoms the word posterior probability. Our experiments have been carried out on the corpus used during ESTER, the french evaluation campaign on automatic transcription of french broadcast news. Using the normalized cross entropy (NCE) as an evaluation metric, experimental results on test data of ESTER evaluation show a very significant improvement : whereas the word posterior probability reaches a value of NCE equal to 0.187, the WP/LMBB measure obtains 0.270.

1 Introduction

Les mesures de confiance servent à estimer la fiabilité d'une hypothèse de reconnaissance et donc, la fiabilité d'un système de traitement de la parole [7,11,12]. Les mesures de confiances sont utilisées en traitement de la parole dans divers champs d'applications. Par exemple, elles peuvent permettre d'extraire des annotations pertinentes de transcriptions automatiques ou encore d'améliorer l'efficacité des systèmes de dialogue grâce à la détection d'erreurs et des mots hors-vocabulaire.
Les mesures de confiance sont généralement estimées grâce aux probabilités *a posteriori* données par le système de reconnaissance. Par exemple, dans [11], les auteurs ont recours aux graphes de mots et aux listes des N meilleures hypothèses pour définir la probabilité *a posteriori* d'un mot comme mesure de confiance. Un inconvénient de la probabilité *a posteriori* d'un mot est la forte sensibilité de cette mesure à la topologie de l'espace de recherche sur lequel elle est calculée (profondeur d'un graphe de mots ou taille d'une liste de N meilleures hypothèses). Ainsi, les différentes heuristiques de réduction de l'espace de recherche utilisées lors du processus de reconnaissance ont une incidence directe sur la valeur de la probabilité *a posteriori* et peuvent altérer sa pertinence.

Dans cet article, nous proposons d'améliorer les capacités de discrimination de la probabilité *a posteriori* en la combinant avec une mesure qui n'est pas affectée par ce type de problème. Cette nouvelle mesure est basée sur le comportement du repli du modèle de langage. Nous verrons que fusionner les deux mesures semble réduire l'impact de l'espace de recherche sur la probabilité *a posteriori* et améliore sa capacité à discriminer une hypothèse correcte d'une hypothèse incorrecte.

2 Mesures de confiance

Soit l'ensemble de mots $\{w_1, \ldots, w_N\}$ composé de N éléments. Chaque mot w peut être associé à une mesure de confiance $m(w)$, qui doit appartenir à l'intervalle $[0,1]$ et doit correspondre à la probabilité que le mot w soit correct. Soit $\mu(m) = \frac{1}{N}\sum_{i=1}^{N} m(w_i)$, la dernière propriété nous donne que, pour une mesure de confiance idéale, $\mu(m)$ doit être une approximation de p_{ok} où p_{ok} est le taux de mots émis bien reconnus (par rapport au WER, le taux de mots émis bien reconnus ne prend pas en compte les suppressions car il s'applique uniquement aux mots émis par le système de transcription).
Comparer la valeur prédictive $\mu(m)$ du taux de mots bien reconnus à la valeur réelle de ce taux peut être une bonne métrique pour évaluer la qualité d'une mesure de confiance. Mais cette métrique permet seulement l'évaluation de la capacité de prédiction globale de la mesure : elle ne reflète pas la pertinence locale de la mesure sur le mot. Cette information locale peut être évaluée grâce à une métrique utilisée lors des campagnes d'évaluation NIST. Cette métrique, appelée entropie croisée normalisée (NCE) sert à évaluer la pertinence des mesures de confiance, c'est une estimation de l'apport en information additionnelle d'une mesure de confiance [3,5] :

$$\begin{aligned} NCE = \{ & H_{max} + \sum_{correctw} \log_2(m(W)) \\ & + \sum_{incorrectW} log_2(1 - m(W)) \} / H_{max} \end{aligned} \quad (1)$$

où $H_{max} = -n\log_2(p_c) - (N-n)\log_2(1-p_c)$,
n est le nombre de mots reconnus corrects,
N , le nombre total de mots reconnus, et
p_c, la probabilité moyenne qu'un mot reconnu soit correct ($= n/N$).

Une mesure de confiance doit donc respecter deux propriétés qui peuvent permettre de l'évaluer :
1. elle donne une prédiction globale sur la probabilté moyenne que le mot soit correct ;
2. elle donne une information locale sur la fiabilité d'un mot hypothèse.

2.1 *Probabilité a posteriori*

Les probabilités *a posteriori* peuvent être calculées à partir des listes des N meilleurs hypothèses [9], des graphes de mots [3] ou des réseaux de confusions [6]. En fait, la probabilité *a posteriori* d'un mot est le taux de la probabilité *a priori* d'un mot sur la somme des probabilités *a priori* de toutes les autres hypothèses alternatives. Ces probabilités *a priori* sont une combinaison des scores fournis par les modèles acoustiques et linguistique.
Dans les listes des N meilleurs hypothèses la probabilité *a posteriori* d'un mot est calculée avec le taux de la somme des probabilités *a priori* des occurrences de ce mot à une position donnée parmi les N hypothèses, sur la somme de toutes les probabilités *a priori* des mots situés à la même position, incluant celles des occurrences du mot courant.
Dans les aproches basées sur les graphes de mots ou les réseaux de confusions, la probabilité *a posteriori* est la généralisation de l'approche précédente où la segmentation en mots et la profondeur de l'espace de recherche sont mieux pris en considération.
Comme cette mesure est dépendante des heuristiques d'élaguage des graphes de mots générés lors du processus de reconnaissance, les scores de confiance obtenus grâce à celle-ci peuvent être biaisés. Pour remédier à cela, [3] utilise un arbre de décision pour transormer les probabilités *a posteriori* des mots en véritables scores de confiance.

Ici, nous utilisons un réseau de confusion basé sur la technique utilisée dans [6] pour calculer les probabilités *a posteriori* des mots. Pour éviter le biais dû aux heuristiques d'élagage, nous combinons la probabilité *a posteriori* avec une autre mesure de confiance qui n'est pas affectée par la taille de l'espace de recherche. La mesure de confiance basée sur les probabilités *a posteriori* des mots sera notée WP.

2.2 *Mesure de confiance linguistique basée sur le comportement du repli*

La mesure de confiance que nous introduisons ici est basée sur le comportement du mécanisme de repli d'un modèle de langage n-gramme, comme dans [10]. Nous appelons $LMBB$ (Language Model Back-off Behavior) cette méthode.
Cette mesure, à partir du modèle de langage utilisé pour la reconnaissance, prend en compte l'ordre du n-gramme le plus élevé qui peut être associé au mot visé par la mesure de confiance et à l'historique de taille $n-1$ de ce mot. Par exemple, si la séquence de mots "... il est temps de ..." est reconnue en utilisant un modèle de langage quadrigramme et que le quadrigramme [il est temps de] a été observé dans le corpus d'apprentissage du modèle de langage, alors le mot 'de' sera associé à l'ordre 4. Par contre, si ce quadrigramme n'a pas été observé, mais que le trigramme [est temps de] l'a été, alors le mot 'de' sera associé à l'ordre 3. De la même manière, ce mot pourrait être associé à l'ordre 2 ou à l'ordre 1 le cas échéant, et même à l'ordre 0 dans le cas peu courant où les mots hors-vocabulaire peuvent être traités.
Un phénomène bien connu en reconnaissance de la parole est la propagation des erreurs : lorsqu'un mot est mal reconnu, les mots qui l'entourent sont souvent également affectés par des erreurs. Dès lors, il semble très intéressant d'intégrer dans la mesure de confiance linguistique d'un mot des informations concernant son voisinage. En supposant que le comportement du mécanisme de repli d'un modèle de langage est un bon indicateur de la fiabilité de ce modèle, nous proposons de prendre également en compte l'ordre associé aux deux mots voisins du mot visé (le voisin de gauche et celui de droite). Dès lors, chaque mot reconnu est associé à trois valeurs d'ordre de n-gramme (l'ordre du n-gramme le plus élevé pouvant être associé au voisin de gauche, celui du voisin de droite, et celui du mot visé lui-même). Ces trois valeurs définissent un nombre finis d'étiquettes, chacune de ces étiquettes représentant une classe de comportement du mécanisme de repli d'un modèle de langage n-gramme. Nous supposons que les mots reconnus appartenant à une même classe de comportement de ce mécanisme sont associés à la même valeur de mesure de confiance linguistique qu'il faut estimer *a priori* pour chacune de ces classes.
Afin de ne pas distinguer un nombre de classes différentes trop important qui seraient difficiles à bien modéliser sans grande quantité de données d'apprentissage, nous ne prendrons pas les valeurs réelles des ordres associés aux mots voisins mais leur position relative par rapport à l'ordre associé au mot visé : plus grand (+), plus petit (-) ou égal (=). Ceci permet de réduire le nombre de classes possibles.

Pour illustrer ce propos, prenons par exemple la séquence de mots "... il est temps de lire ce livre..." et supposons :
- que le quadrigramme [il est temps de] et le trigramme [est temps de] n'ont pas été observés dans le corpus d'apprentissage du modèle de langage, alors que le bigramme [temps de] l'a été : le mot 'de' est associé à l'ordre 2 ;
- que le quadrigramme [est temps de lire] n'a pas été observé dans le corpus d'apprentissage alors que le trigramme [temps de lire] l'a été : le mot 'lire' est associé à l'ordre 3 ;
- que le quadrigramme [temps de lire ce] a été observé dans le corpus d'apprentissage : le mot 'ce' est associé à l'ordre 4.

Dès lors, la classe de comportement du mot 'lire' sera associée à l'étiquette (-,3,+), car le mot 'lire' est associé à l'ordre 3, son voisin de gauche est associé à un ordre inférieur (-) de valeur 2 et son voisin de droite est associé à un ordre supérieur (+) de valeur 4.
En comparant un ensemble de transcriptions automatiques dont les mots sont marqués par ce type d'étiquettes, et en ayant pour les enregistrements audio de ces transcriptions des transcriptions manuelles, il est aisé de calculer le taux d'erreur de reconnaissance pour les mots qui composent chacune de ces classes. Ce taux d'erreur est le rapport entre le nombre de mots $n_{err}(cl)$ mal reconnus (substitutions ou insertions) contenus dans une classe cl sur le nombre de mots $n_{mots}(cl)$ qui composent cette classe (pour un ensemble de transcriptions donné). Ainsi, pour un mot w associé à la classe cl, la valeur $m_{lmbb}(w)$ donnée par la mesure de confiance LMBB se calcule à partir d'un corpus d'apprentissage composé de trans-

criptions automatiques et manuelles avec la formule :

$$m_{lmbb}(w) = 1 - \frac{n_{err}(cl)}{n_{mots}(cl)}$$

La figure 1 montre qu'il existe véritablement une corrélation entre le comportement du mécanisme de repli du modèle de langage et le taux d'erreur. Ces résultats ont été calculés sur le corpus d'apprentissage décrit section 3.2.

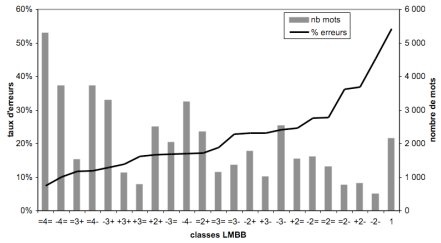

Fig. 1: Taux d'erreur, répartition des mots transcrits et classes LMBB

2.3 Fusion WP/LMBB

Fusionner la probabilité *a posteriori* avec une autre mesure qui n'est pas affectée par la taille de l'espace de recherche et qui est indépendante de celle-ci devrait améliorer ses résultats en tant que mesure de confiance. En effet, même si la mesure LMBB est dérivée du modèle de langage utilisé lors du décodage, ce n'est pas le score de celui-ci qui est pris en considération. Cette mesure apporte donc d'autres informations qui ne sont pas prises en compte lors du calcul de la probabilité *a posteriori*.

Combiner plusieurs paramètres pour obtenir une mesure unique peut être fait de plusieurs façons [8]. Les opérateurs les plus utilisés dans cette fusion sont : minimum, maximum, moyenne arithmétique, moyenne géométrique, produit ou encore la moyenne quadratique. Pour prendre en compte les qualités de chacune des mesures, une moyenne pondérée $m(w) = \sum_{i=1}^{K} q_k m_k(w)$, avec $\sum_{k=1}^{K} q_k = 1$ peut être utilisée. Les poids q_k peuvent être appris empiriquement par validation croisée (voir résultats de la section 2). Diverses techniques de fusion provenant de théories telles que la théorie de l'évidence ou encore la théorie des probabilités ont été essayées mais la technique qui a donné les meilleurs résultats sur le corpus d'apprentissage des mesures de confiance (CTrain) est une simple interpolation linéaire. La mesure retenue est notée WP/LMBB.

3 Expériences et résultats

Les différentes expériences décrites dans cet article sont basées sur le corpus d'ESTER, campagne d'évaluation sur des systèmes de transcriptions d'émissions radiophoniques en français démarrée en 2003 et achevée en janvier 2005 [4].
Le système utilisé durant cette campagne par le Laboratoire d'Informatique de l'Université du Maine (LIUM) est basé sur le décodeur CMU Sphinx 3.3. Plusieurs paramètres ont été ajoutés, comme l'adaptation de modèles acoustiques utilisant la méthode SAT (Speaker Adaptative Training) ou encore le rescoring de graphes de mots utilisant des modèles de language quadrigrammes. Ce système a atteint la seconde position de la campagne avec 23.6% de taux d'erreur mot –incluant les insertions, substitutions et suppressions [1, 2].

3.1 Modèles acoustiques et linguistique

Le dictionnaire utilisé contient environ 65K mots. Les modèles acoustiques ont été appris avec 81h de transcriptions manuelles provenant de différentes radios : France Inter, France Info, Radio Télévision Marocaine (RTM) et Radio France International (RFI). Ce sont des émissions d'actualités. Ces émissions sont majoritairement en bande large mais comportent également de de la bande étroite (téléphone). Pour cet apprentissage, nous avons utilisé le toolkit SphinxTrain, associé aux décodeurs de Sphinx CMU. Les modèles acoustiques sont appris avec une différenciation bande large/bande étroite. Le corpus de développement pour réévaluer les différents paramètres est constitué de 4h provenant de ces différentes radios. Le modèle de langage trigramme est utilisé lors des deux premières phases de décodage. La troisième phase utilise un modèle quadrigramme qui correspond à un rescoring de graphes de mots.

3.2 Apprentissage des paramètres pour les mesures de confiance

Les mesures de confiance ont été élaborées grâce à un échantillon de 4h provenant de France Inter, France Info, RTM et RFI. Ces 4h sont indépendantes du corpus d'apprentissage des modèles acoustiques et linguistique et sont fournies avec leur transcription manuelle. Une transcription automatique est obtenue avec le système de reconnaissance. Á partir des transcriptions automatique et manuelle, nous avons pu calculer les scores de confiance de la mesure LMBB à partir du taux d'erreur obtenu avec les différentes classes de LMBB (voir figure 1). Les probabilités *a posteriori* des mots sont calculées sur les graphes fournis par le système. Plusieurs mesures fusionnées ont été comparées sur CTrain et la meilleure en termes de NCE est : $m_{WP/LMBB}(w) = 0.7 * m_{WP}(w) + 0.3 * m_{LMBB}(w)$.

3.3 Résultats

Le corpus de test officiel (noté CTest) est composé de 10h d'émissions de radio (environ 10 000 phrases et 114 000 mots) : 2h de chacune des 4 radios du corpus d'apprentissage ainsi que 2h provenant de deux radios inconnues au moment de l'évaluation.
Nous avons vu à la section 2 que la moyenne des scores donnée par une mesure doit approximer le taux de mots émis bien reconnus. Le tableau 1 montre que la probabilité *a posteriori* ne permet pas d'obtenir ce taux sur les données de test. C'était aussi le cas sur CTrain. Pour pallier ce problème, une méthode classique de normalisation par transformation sigmoïdienne a été calculée mais cette méthode ne donnait pas d'amélioration en termes de NCE. Par contre, le taux de prédiction globale de la mesure LMBB est proche du taux de mots émis corrects. Ceci s'explique par le fait que les corpus CTest et CTrain sont proches en termes de couverture du modèle de langage. De

plus, les classes LMBB ont été apprises sur CTrain et sont donc bien représentatives de l'impact du comportement du repli du modèle de langage. La mesure fusionnée WP/LMBB propose donc un taux de prédiction globale amélioré par rapport à celui de la probabilité *a posteriori* seule.
De plus, nous avons pu remarqué que beaucoup de mots ont une probabilité supérieure à 0,9 et pourtant, ils ne sont pas corrects. Une explication possible est que, dans l'espace de recherche, ces mots n'ont pas ou peu d'alternative. Néanmoins, ces mots ne sont pas pertinents du point de vue de la mesure LMBB et ont donc un score LMBB faible. La mesure LMBB permet donc à la probabilité *a posteriori* d'être plus discriminante.
Enfin, pour évaluer les mesure localement sur les deux corpus grâce à la NCE, le tableau 2 montre que la probabilité *a posteriori* donne une réelle information sur l'exactitude d'un mot d'une manière plus significative que la mesure LMBB. La mesure WP/LMBB obtenue en fusionnant les deux mesures améliore nettement les résultats de la probabilité *a posteriori* prise seule car les scores augmentent de 0,187 à 0,270 sur CTest. Ceci est dû au fait que les deux mesures initiales apportent des informations complémentaires.

Tab. 1: Taux de prédiction globale des mesures de confiance sur les données de test

Mesure	Prédiction globale
LMBB	83.3%
probabilité *a posteriori*	72.6%
WP/LMBB	75.2%
Taux correct réel	***80.8***%

Tab. 2: Comparaison des diverses mesures de confiance sur les données d'apprentissage et de test des mesures de confiance en termes d'entropie croisée normalisée(NCE)

Mesure	CTrain	CTest
LMBB	0.081	0.063
probabilité *a posteriori*	0.169	0.187
WP/LMBB	**0.278**	**0.270**

4 Conclusion

Dans cet article, nous proposons une amélioration de la probabilité *a posteriori* prise comme mesure de confiance. Pour obtenir une meilleure mesure, nous la fusionnons avec une autre mesure provenant d'une autre partie du système de reconnaissance pour bénéficier d'informations complémentaires à apporter à la probabilité *a posteriori*. Cette mesure est calculée à partir du comportement du repli du modèle de langage et est négligeable en termes de temps de calcul. Leur fusion, WP/LMBB, est une simple interpolation linéaire des deux mesures et améliore nettement leurs capacités à prédire si le mot émi est correct ou non.
Cette mesure de confiance est pertinente pour de multiples applications du traitement de la parole. Avec une telle mesure, les mots fiables sont plus aisément détectables pour exploiter les sorties du système de reconnaissance, ou encore pour vérifier les capacités du système sur des tâches spécifiques. Par la suite, il serait intéressant d'essayer d'autres mesures pertinentes à fusionner ainsi que d'autres méthodes de fusion comme les arbres de décisions. Enfin, il serait intéressant d'étudier la mesure WP/LMBB dans des applications du traitement de la parole telle que le processus de décodage.

Références

[1] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[2] Y. Estève, P. Deléglise, and B. Jacob. Système de transcription automatique de la parole et logiciels libres. *Traitement Automatique Des Langues*, 45(2), 2004.

[3] G. Evermann and P.C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Speech Transcription Workshop*, 2000.

[4] G. Gravier, J.-F. Bonastre, S. Galliano, and E. Geoffrois. The ESTER evaluation campaign of rich transcription of french broadcast news. In *LREC, Language Evaluation and Resources Conference*, Lisbon, Portugal, May 2004.

[5] B. Maison and R. Gopinath. Robust confidence annotation and rejection for continuous speech recognition. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001.

[6] H. Mangu, E. Brill, and Stolcke A. Finding consensus in speech recognition : Word error minimization and other applications of confusion networks. *Computer Speech and Language*, pages 4373–400, 2000.

[7] R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, and J. Pardo. Confidence measures for spoken dialogue systems. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001.

[8] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 875–878, Munich, Allemagne, April 1997.

[9] A. Stolcke, Y. Konig, and M. Weintraub. Explicit word error minimization in N-best list rescoring. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, volume 1, pages 163–166, Rhodes, Greece, 1997.

[10] C. Uhrick and W. Ward. Confidence metrics based on n-gram language model backoff behaviors. In *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Rhodes, Greece, September 1997.

[11] F. Wessel and H. Ney. Unsupervised training of acoustic models for large vocabulary continous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13 :23–31, 2005.

[12] G. Williams. A study of the use and evaluation of confidence measures in automatic speech recognition. Technical report, Department of Computer Science, University of Sheffield, March 1998.