Mesures de confiance trame-synchrone

Joseph Razik, Odile Mella, Dominique Fohr et Jean-Paul Haton

Loria - UMR 7503 - Equipe Parole
Vandœuvre-Lès-Nancy, France
prenom.nom@loria.fr http ://parole.loria.fr

ABSTRACT

This paper presents some confidence measures for large vocabulary speech recognition that can be evaluated directly within the first steps of the recognition process. Having some clues to drive the recognition process may help to improve the accuracy of a sentence. Confidence measures may fit to this goal, so we propose some measures that can help the recognition as early as possible, without having to wait for the recognition process to be completed. Furthermore, our confidence measures are local, and they are based on partial word graphs. Experiments on a French broadcast news corpus are presented, and give results close to the post calculated version of the measures.

1. INTRODUCTION

Dans la plupart des systèmes de reconnaissance automatique de la parole, l'utilisation d'une mesure de confiance associée à un mot reconnu se fait lors d'une étape postérieure distincte. Ceci est principalement dû à leur objectif : validation dans un processus de recherche de mots clés [2], sélection de mots corrects en apprentissage non supervisé [8] ou bien encore détection de mots hors vocabulaire [3] .

Or, il serait parfois intéressant d'avoir accès à une mesure de confiance directement pendant le processus de reconnaissance et donc au sein même du moteur de reconnaissance. Ainsi, le moteur pourrait tenir compte d'un effet de *doute* sur des mots et modifier le processus de reconnaissance en conséquence.

Nous nous plaçons dans ce cadre en proposant dans cet article des mesures intégrables directement dans le moteur de reconnaissance et calculables quasiment de manière synchrone avec la progression de la reconnaissance. En effet, la plupart des mesures nécessitent le décodage complet de la phrase à reconnaître pour prendre une décision ou donner une estimation. Nous utilisons comme moteur de reconnaissance Julius [5], un moteur grand vocabulaire effectuant un traitement en deux passes. La première passe détermine pour chaque trame un nombre restreint de mots hypothèses. La deuxième passe fournit la phrase finale ayant la plus grande vraisemblance. Les mesures que nous proposons se placent au niveau du mot ; elles ont un caractère local et interviennent directement au cours du déroulement de la première passe.

Plusieurs approches ont été proposées afin de calculer des mesures de confiance, comme la mise en compétition de modèles, une comparaison avec le résultat d'un décodeur phonétique [4] ou l'utilisation de paramètres heuristiques (nombre de phonèmes dans le mot, durée des phonèmes, etc.). D'autres méthodes tentent d'estimer la probabilité *a posteriori* des mots reconnus ou bien sont basées sur les informations issues d'un graphe de mots contenant les multiples chemins qui vont du début à la fin de la phrase (Ortmanns [6]). Les mesures présentées dans cet article font partie de cette dernière approche. Plus précisément, elles sont inspirées de mesures à caractère local proposées dans nos travaux précédents [7], et d'une mesure fondée sur la probabilité *a posteriori* et décrite par F. Wessel et al. [10].

La section 2 décrit les mesures de confiance développées (mesures locales et mesure de la probabilité *a posteriori*). La section 3 décrit les conditions d'expérimentation ; elle contient une description rapide du moteur de reconnaissance, de sa structure interne (dans laquelle nous puisons les informations pour les différentes mesures) et des corpus qui sont utilisés. Pour finir, les différents résultats et tests sont décrits en section 4.

2. MESURES DE CONFIANCE

Le but d'une mesure de confiance est en général de donner une estimation de la probabilité qu'un mot reconnu soit correct. Les mesures décrites dans cet article permettent de donner cette estimation directement au cours du processus de reconnaissance du moteur, pendant le déroulement de la première passe de celui-ci. Pendant cette passe, le moteur génère une structure interne contenant pour chaque trame un nombre restreint d'hypothèses de mots possibles. Cette structure servira au cours de la deuxième passe à déterminer la phrase la plus vraisemblable. Afin de pouvoir fournir une mesure de confiance pour les mots pendant le déroulement de la première passe, les mesures présentées ont un caractère local : elles n'utilisent que des informations disponibles au moment du calcul. Nous présentons dans cette section des mesures de confiances locales. Toutes sont basées sur le graphe de mots interne au moteur de reconnaissance, tandis qu'une seule est issue de la mesure décrite par F. Wessel et al. [10] estimant la probabilité *a posteriori*.

Introduisons quelques notations : soit w un mot hypothèse, τ son instant de début et t son instant de fin. Une phrase commence au temps 1, se termine au temps T et x_1^T représente la séquence d'observations du temps 1 au temps T. Soit $[w, \tau, t]$ un mot hypothèse spécifique, et $[w, \tau, t]_1^M$ une séquence de M mots $[w_i, \tau_i, t_i]$, où $\tau_1 = 1, t_M = T$ et $t_{i-1} = \tau_i - 1$ pour $i = 2, \ldots, M$. $C([w, \tau, t])$ représente la mesure de confiance pour le mot hypothèse $[w, \tau, t]$.

2.1. *Mesures de confiance locales*

Pour concevoir nos mesures de confiance, nous utilisons un graphe de mots extrait du treillis d'exploration du moteur de reconnaissance. L'idée est de pouvoir calculer pendant la phase de reconnaissance du moteur une mesure de confiance pour chaque mot de la phrase.

Mesure basée sur des probabilités unigrammes Cette mesure utilise seulement des informations très simples et très locales : les scores acoustiques et les probabilités unigrammes. Cette mesure est similaire à un rapport de vraisemblance [9] entre le mot analysé et d'autres mots hypothèses, mais nous ne prenons en compte dans le rapport que les mots ayant survécu à l'élagage du faisceau de recherche et satisfaisant des conditions sur leur temps de début, de fin, et sur leur durée. Pour cette mesure, nous utilisons le graphe de mots interne du moteur de reconnaissance afin de sélectionner les mots hypothèses. Pour cela, nous introduisons un facteur de relâchement sur les contraintes temporelles de sélection de mots dans le graphe. Ces contraintes concernent les temps de début et de fin, et la longueur des mots hypothèses. Par exemple, pour un mot hypothèse $[w, \tau, t]$ et un taux de relâchement de 0,5, nous considérons les mots qui apparaissent avec un temps de début égal à $\tau \pm 50\%$ de la longueur de w. Le temps de fin et la longueur du mot sont traités de la même manière. Or, avec ce relâchement de contraintes, plusieurs occurrences du même mot hypothèse satisfaisant ces nouvelles contraintes peuvent apparaître. Dans ce cas, l'hypothèse ayant obtenu le score acoustique maximal est retenue. Nous introduisons également des facteurs d'échelle, à la fois pour le score acoustique (α) et pour le score du modèle de langage (β). Notre première mesure est ainsi définie par l'équation Eq. 1.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha . p(w)^\beta}{\sum_{[w',\tau',t'] \in E} max(p(x_{\tau'}^{t'}|w'))^\alpha . p(w')^\beta} \quad (1)$$

où E est l'ensemble des mots qui satisfont les contraintes de temps et de longueur données par le facteur de relâchement.

Mesures basées sur des probabilités bigrammes La mesure précédente est modifiée afin de prendre en compte des informations sur le voisinage du mot par l'intermédiaire de probabilités bigrammes. Ces probabilités pour un mot $[w, \tau, t]$ sont calculées avec tous les mots précédents w_p qui se terminent au temps $\tau - 1$. Nous obtenons ainsi l'équation Eq. 2.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha \sum_{w_p} \left(p(w|w_p)p(w_p)\right)^\beta}{\sum_{[w',\tau',t'] \in E} max(p(x_{\tau'}^{t'}|w'))^\alpha \sum_{w'_p} \left(p(w'|w'_p)p(w'_p)\right)^\beta} \quad (2)$$

Les informations apportées par la probabilité bigramme peuvent ne pas être suffisantes car encore trop locales. C'est pourquoi nous collectons encore un peu plus d'informations en utilisant les probabilités bigrammes avec les mots *précédents* et *suivants*. Pour un mot hypothèse $[w, \tau, t]$, nous considérons tous les mots précédents possibles w_p qui se terminent à $\tau - 1$, et tous les mots suivants possibles w_s qui commencent à $t + 1$. Nous introduisons une notation supplémentaire Γ, qui représente pour un mot hypothèse $[w, \tau, t]$ les informations issues des probabilités bigrammes :

$$\Gamma_{[w,\tau,t]} = \sum_{w_p} \sum_{w_s} \{p(w|w_p).p(w_s|w).p(w_p)\}^\beta \quad (3)$$

Nous définissons notre mesure par l'équation Eq. 4.

$$C([w,\tau,t]) = \frac{max(p(x_\tau^t|w))^\alpha \Gamma_{[w,\tau,t]}}{\sum_{[w',\tau',t'] \in E} max(p(x_{\tau'}^{t'}|w'))^\alpha \Gamma_{[w',\tau',t']}} \quad (4)$$

2.2. *Les mesures basées sur la probabilité a posteriori*

La probabilité *a posteriori* est un bon indicateur de l'exactitude d'un mot et beaucoup de mesures de confiance s'appuient sur cette probabilité. F. Wessel et al. [10] ont proposé une mesure de confiance définie par la probabilité *a posteriori* d'un mot. Leur méthode pour calculer cette probabilité est inspirée de l'algorithme *forward-backward*, mais appliqué cette fois avec la granularité du mot. Cet algorithme est appliqué à un graphe de mots semblable à celui généré par le moteur de reconnaissance. La mesure définie par F. Wessel et al. [10] nécessite que le graphe de mots soit totalement généré pour pouvoir déterminer la probabilité *a posteriori*. Ainsi, une utilisation directe pendant le processus de reconnaissance n'est pas possible.

Nous proposons alors de modifier cette mesure afin d'obtenir une mesure plus locale. L'idée est simple : considérer pour chaque mot, non pas le graphe entier, mais un sous graphe contenant le mot à analyser. Pour chaque mot, nous déterminons un voisinage centré sur celui-ci, délimitant une plage temporelle à partir de laquelle nous extrayons un sous-graphe du graphe total. Puis, ce sous graphe est vu comme le graphe de mots associé à une *pseudo phrase* équivalente à une sous-séquence de la phrase. Afin de déterminer ce voisinage, le calcul de la mesure est en retard par rapport à la progression temporelle du moteur de reconnaissance, mais seulement de quelques trames.

Pour ces mesures basées sur la probabilité *a posteriori*, nous avons également utilisé des facteurs d'échelle et un facteur de relâchement. Il est à noter que les mesures sont calculées avec des probabilités bigrammes.

3. CONDITIONS D'EXPÉRIMENTATION

Pour chacune des différentes mesures précédemment définies, les conditions d'expérimentation sont identiques, et ces mesures ont accès exactement aux mêmes données. Nous décrivons dans cette section les différents paramètres définissant ces conditions d'expérimentation.

3.1. *Les modèles acoustiques, de langage et le lexique*

Le corpus d'apprentissage des modèles acoustiques utilisés par le système de reconnaissance se compose de 7 heures de bulletins d'informations radiophoniques, contenant uniquement de la parole large bande (pas de téléphone, pas de musique pure et pas de parole sur fond musical). Le signal est paramétré par des MFCC en appliquant une normalisation MCR (Mean Cepstral Remo-

val). Chaque phonème est modélisé à l'aide d'un modèle HMM.

Le modèle de langage a été appris par l'intermédiaire du CMU Toolkit [1] sur 16 ans du journal français « Le Monde », complété par une transcription manuelle de tout le corpus d'apprentissage de bulletins d'informations radiophoniques. Finalement, nous avons 2.5M de bigrammes et 5.8 M de trigrammes.

Le lexique de 54747 mots contient à la fois des mots au sens habituel du terme, mais aussi des groupes de mots. En effet certains mots ont été regroupés en une seule entité dans le lexique et ne comptent donc que pour un *mot*. Par exemple, la séquence « de la » est représentée par une seule entité « de_la », mais aussi certains noms comme « Aix_les_Bains ».

3.2. Le moteur de reconnaissance Julius

Julius [5] est un système de reconnaissance de la parole grand vocabulaire. Le processus de reconnaissance s'effectue en deux passes : une première passe trame-synchrone qui génère un treillis d'exploration en utilisant un modèle de langage bigramme, et une deuxième passe utilisant ce treillis et des modèles trigrammes pour aboutir à la phrase reconnue. Nous nous servons de Julius dans sa version 3.4.1-multipath, compilée avec l'option v2.1 pour une précision accrue.

Le treillis d'exploration Cette structure interne du moteur de reconnaissance, générée pendant la première passe, donne accès pour chaque trame du signal à plusieurs informations : les hypothèses de mots pouvant se terminer à cette trame, le mot précédent, leur score acoustique et de modèle de langage, etc. Nous obtenons en moyenne 470 mots hypothèses par trame avec un maximum de 2523 mots. En fait, le treillis d'exploration peut être considéré comme un graphe de mots.

3.3. Le corpus de développement et de test

Un corpus, également constitué de bulletins d'informations radiophoniques mais indépendant de celui utilisé pour l'apprentissage des modèles acoustiques, a été divisé en deux parties : une pour le développement, et une pour les tests. Le corpus de développement, d'une durée de 56 minutes, sert à mettre au point le seuil de décision et les facteurs d'échelle des mesures. Le corpus de test est d'une durée de 53 minutes. Ces corpus contiennent respectivement 12135 et 11272 mots. Le taux de reconnaissance moyen sur les deux corpus est d'environ 70,9%. L'ensemble du corpus est constitué de parole large bande, sans parole téléphonique ni musique, mais des phrases peuvent contenir un bruit ou une musique de fond. Le nombre moyen de *mots* par phrase du corpus de test est de 11,5.

3.4. L'évaluation

Pour évaluer les différentes mesures, nous étiquetons les mots de la phrase en deux classes : acceptation et rejet. Cet étiquetage dépend d'un seuil qui définit une frontière entre ces deux classes. Les étiquettes des mots sont ensuite comparées aux fichiers de référence. Ainsi, nous pouvons évaluer deux taux : le taux de *Fausse Acceptation* (FA) et le taux de *Faux Rejet* (FR). Le taux de fausse acceptation correspond aux cas où un mot incorrect est accepté, et le taux de faux rejet correspond aux cas où un bon mot est rejeté.

$$FA = \frac{nb.\ de\ mots\ incorrects\ étiquetés\ Acceptation}{nb.\ de\ mots\ incorrects} \quad (5)$$

$$FR = \frac{nb.\ de\ mots\ corrects\ étiquetés\ Rejet}{nb.\ de\ mots\ corrects} \quad (6)$$

A l'aide de ces deux taux et avec différents seuils de confiance, nous pouvons représenter la courbe DET (Detection-Error Tradeoff) et en déduire le taux EER d'égale erreur (Equal Error Rate).

4. TESTS ET RÉSULTATS

Pour nos tests, nous utilisons plusieurs valeurs pour le facteur de relâchement : 0,1 ; 0,2 ; 0,3 et 0,5. Concernant les facteurs d'échelle, nous considérons deux couples de valeur : $(\alpha;\beta) = (1;1)$ et $(\alpha;\beta) = (0,1;0,95)$. Le deuxième couple (0,1 ; 0,95) correspond aux valeurs optimales pour la mesure de confiance de F. Wessel et al. [10] obtenues sur notre corpus de développement. Des expérimentations supplémentaires montrent que ce couple de valeur est également optimal pour les autres mesures définies. Dans tous les tableaux qui suivent, les valeurs représentent le taux EER.

Dans le premier test, nous considérons notre mesure simple (Eq. 1) qui ne repose que sur les scores acoustiques et les probabilités unigrammes.

TAB. 1: Taux d'EER avec la probabilité unigramme (Eq. 1).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	39,9%	41,5%	43,5%	45,5%
$(0,1;0,95)$	**39,4%**	**39,4%**	40,6%	43,1%

Les résultats de la table 1 montrent principalement l'importance des facteurs d'échelle dans l'amélioration du taux EER pour cette mesure.
Ensuite, nous testons l'influence de l'introduction des probabilités bigrammes avec les mots précédents (Eq. 2). La mesure reste locale et ne dépend que du voisinage passé du mot hypothèse courant.

TAB. 2: Taux d'EER avec la probabilité bigramme arrière (Eq. 2).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	39,7%	39,9%	42,7%	44,7%
$(0,1;0,95)$	38,8%	**38,7%**	39,6%	42,8%

L'introduction des probabilités bigrammes se traduit par une amélioration des résultats (table 2), mais pas de manière importante. C'est pourquoi nous définissons une troisième mesure qui prend en compte un plus grand voisinage du mot hypothèse (Eq. 4). En effet, cette mesure se base sur les probabilités bigrammes avec les mots précédents et suivants. Là encore, nous observons une amélioration (table 3). Moyennant un retard de quelques trames, cette mesure peut encore être évaluée au cours de

la phase de construction du graphe de mots. Ce délai est nécessaire à l'obtention d'une stabilité dans le graphe pour la sélection des mots suivants.

TAB. 3: Taux d'EER avec les probabilités bigrammes avant et arrière (Eq. 4).

	Taux de relâchement			
$(\alpha;\beta)$	0,1	0,2	0,3	0,5
$(1;1)$	42,5%	40,8%	39,7%	39,8%
$(0,1;0,95)$	39,8%	37,7%	**36,8%**	39,3%

La table 4 présente les résultats de notre dernière mesure de confiance à caractère local fondée sur la probabilité *a posteriori* locale et utilisant un voisinage plus important. Les résultats sont présentés selon la taille en mots de la *pseudo phrase* centrée sur le mot hypothèse courant. Pour déterminer la taille du voisinage en nombre de mots, nous nous basons sur la meilleure hypothèse de phrase. Avec un faible retard par rapport à la progression du moteur de reconnaissance, nous pouvons déterminer les mots qui précèdent et suivent le mot analysé.

TAB. 4: Taux d'EER avec la probabilité *a posteriori* locale.

		Taux de relâchement	
$(\alpha;\beta)$	Nb. mots	0,2	0,5
$(1;1)$	1	40,7%	40,8%
	3	37,1%	37,1%
	5	36,0%	36,0%
	7	35,6%	35,7%
$(0,1;0,95)$	1	43,4%	43,4%
	3	31,4%	31,5%
	5	**25,6%**	**25,6%**
	7	**24,3%**	**24,3%**

Nous pouvons remarquer l'influence de la longueur de la pseudo phrase. Plus on se rapproche de la longueur de la phrase complète, plus les résultats s'améliorent. Nous avons choisi comme mesure de référence la mesure de F. Wessel et al. [10] : mesure de la probabilité *a posteriori* calculée sur toute la phrase. A partir d'une pseudo phrase de 5 mots, la mesure locale donne des résultats proches de ceux de la mesure de référence (table 5). L'influence du voisinage diminue au delà d'une pseudo phrase de 5 mots, la longueur moyenne des phrases étant de 11,5 mots. Ce phénomène est sans doute dû à l'utilisation d'un modèle de langage bigramme. Ce phénomène reflète cependant l'aspect réel de la construction classique d'une phrase. Nous avons également testé cette mesure en définissant la taille du voisinage en nombre de trames et en extrayant le sous-graphe correspondant. Pour une taille de voisinage de 84 trames (longueur moyenne de 2 mots) de part et d'autre du mot analysé, nous obtenons le même résultat.

TAB. 5: Taux d'EER avec la probabilité *a posteriori* globale.

	Taux de relâchement	
$(\alpha;\beta)$	0,2	0,5
$(1;1)$	35,2%	35,1%
$(0,1;0,95)$	25,4%	**23,8%**

5. CONCLUSION

Dans cet article, nous avons présenté plusieurs mesures de confiance qui répondent à une contrainte forte : la mesure doit être utilisable pendant le processus de décodage du moteur de reconnaissance. Cela signifie que ces mesures ne nécessitent pas l'exécution complète du processus de reconnaissance pour être calculées. Au pire, certaines nécessitent un léger délai par rapport à la progression du moteur. Plus la mesure prend en compte d'information sur son voisinage proche, plus elle est pertinente. Une pseudo phrase de 5 mots est un bon compromis entre taux d'EER et délai pour pouvoir effectuer le calcul de la mesure. Notre meilleur taux d'EER, 24,3%, est atteint avec la mesure de la probabilité *a posteriori* locale et une pseudo phrase de 5 mots. Avec cette mesure, nous respectons notre objectif de réaliser une mesure trame-synchrone. Le taux d'EER obtenu est proche de celui de la mesure de référence, 23,8%, de F. Wessel et al. [10]. Ainsi, nous avons proposé des mesures de confiance pouvant être utilisées directement au cours du processus de reconnaissance ou bien encore à la demande pour valider un mot. Une continuation logique de ce travail consisterait à modifier automatiquement le processus de reconnaissance afin d'améliorer le taux du système.

RÉFÉRENCES

[1] P.R. Clarckson and R. Resenfeld. Statistical language modelling using the CMU-Cambridge toolkit. In *Eurospeech, Rhodes*, pages 2707–2710, 1997.

[2] L. Ferrer and C. Estienne. Improving performance of a keyword spotting system by using a new confidence measure. In *Eurospeech, Aalborg*, pages 2561–2564, 2001.

[3] T. Jitsuhiro, S. Takahashi, and K. Aikawa. Rejection of out-of-vocabulary words using phoneme confidence likelihood. In *ICASSP, Seattle*, pages 217–220, 1998.

[4] S.O. Kamppari and T.J. Hazen. Word and phone level acoustic confidence scoring. In *ICASSP, Istanbul*, 2000.

[5] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. In *Eurospeech, Aalborg*, pages 1691–1694, 2001.

[6] S. Ortmanns and H. Ney. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech and Language*, 11 :43–72, 1997.

[7] J. Razik, O. Mella, D. Fohr, and J.P. Haton. Local word confidence measure using word graph and n-best list. In *INTERSPEECH, Lisbon*, pages 3369–3372, 2005.

[8] F. Wallhoff, D. Willett, and G. Rigoll. Frame-dicriminative and confidence-driven adaptation for LVCSR. In *ICASSP, Istanbul*, pages 1835–1838, 2000.

[9] M. Weintraub. LVCSR log-likelihood ratio scoring for keyword spotting. In *ICASSP, Detroit*, pages 297–300, 1995.

[10] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Proc.*, 9 :288–298, 2001.