Mesures de confiance locales et trame-synchrones

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

LORIA-INRIA – Equipe Parole
Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
http ://parole.loria.fr

ABSTRACT

This paper presents several new confidence measures with the major advantage that they can be evaluated as soon as possible without having to wait for the recognition process to be completed : synchronously with the frame processed by the engine or with a slight delay.
Such measures are useful to drive the recognition process by modifying the likelihood score or to validate recognized words in on-the-fly applications as keyword spotting task and on-line automatic speech transcription for deaf people.
The EER evaluation on a French broadcast news corpus shows performance close to the batch version of these measures (23.0% against 22.0% of EER) with only 0.84s of data before and after the word to analyze.

Keywords : speech recognition, confidence measures, likelihood ratio, posterior probabilities frame-synchronous, local.

1. INTRODUCTION

Pour de nombreux problèmes il est nécessaire de prendre une décision à partir d'un résultat déterminé. Or il est important de savoir quel niveau de confiance il est possible d'accorder à ce résultat. La définition de mesures de confiance est un moyen de répondre à cette problématique : estimer la probabilité qu'un résultat soit correct. En reconnaissance automatique de la parole, la mesure de confiance d'un mot permet d'estimer si le mot reconnu par le moteur de reconnaissance est correct ou non et peut être utile dans des applications comme l'apprentissage non-supervisé, la détection de mots hors-vocabulaire ou la détection de mots clés.
Nous nous sommes intéressés à la définition de mesures de confiance utilisables dans des applications en flux (par exemple la transcription d'émissions en direct ou de cours dans une salle de classe normale pour les malentendants), c'est-à-dire à des mesures calculables le plus tôt possible dans le processus de reconnaissance et sans avoir à attendre la reconnaissance de la totalité de la phrase.
Dans la littérature, plusieurs critères ont été proposés afin de calculer des mesures de confiance : heuristiques [6], linguistiques [5], probabilistes [1, 7]. Cependant, la plupart de ces mesures, notamment les plus précises d'entre elles comme l'estimation de la probabilité *a posteriori*, nécessitent la reconnaissance complète de la phrase prononcée voire plusieurs reconnaissances successives ou bien ne sont applicables qu'à un vocabulaire restreint.
Nous avons donc défini deux types de mesures de confiance, dites trame-synchrones et locales, utilisables par des applications en flux nécessitant une reconnaissance grand vocabulaire. Cet article présente tout d'abord nos mesures de confiance, puis décrit les conditions d'expérimentation, avant de détailler les résultats de leur évaluation sur un corpus radiophonique.

2. MESURES DE CONFIANCE

2.1. *Mesures trame-synchrones*

Les mesures de confiance trame-synchrones utilisent uniquement les informations disponibles à un instant donné de la progression du moteur de reconnaissance. Ainsi, dès qu'une trame du signal est traitée par le moteur, une valeur de confiance peut être estimée pour un mot finissant à cette trame.
Soit $[w, \tau, t]$ un mot débutant à l'instant τ et se terminant à l'instant t, nos mesures de confiance sont fondées sur un rapport de vraisemblance entre le modèle du mot à tester et un ensemble E de modèles compétitifs. Les mesures sont définies selon l'équation générique suivante :

$$C([w,\tau,t]) = \frac{P(O|w)P(w)}{\sum\limits_{w' \in E} P(O|w')P(w')}$$

$P(O|w)$ est la probabilité acoustique associée au mot w et à la séquence d'observation O, $P(w)$ est la probabilité linguistique de w.
Nous avons choisi pour E l'ensemble des mots en concurrence avec w lors de la phase de décodage du moteur de reconnaissance et ayant une position temporelle proche de celle de w. Pour cela nous avons introduit un facteur de relâchement ε tel qu'un mot $[w', \tau', t']$ de E respecte les contraintes suivantes : $\tau - \varepsilon\, d \leq \tau' \leq \tau + \varepsilon\, d$; $t - \varepsilon\, d \leq t' \leq t$; $(1-\varepsilon)\, d \leq d' \leq (1+\varepsilon)\, d$, avec d, durée du mot $[w, \tau, t]$.
Les différentes mesures de confiance trame-synchrones que nous avons définies se distinguent par l'ordre du modèle utilisé pour la probabilité linguistique : unigramme, bigramme ou trigramme mais également par le sens du modèle de langage (direct ou inverse), le choix des mots précédents w et la prise en compte des occurrences multiples d'un mot [4]. Dans le cas bigramme, la mesure de confiance est définie par l'équation suivante :

$$C([w,\tau,t]) = \frac{p(o_\tau^t|w)^\alpha \sum\limits_{w_p} \left(p(w|w_p)p(w_p)\right)^\beta}{\sum\limits_{[w',\tau',t'] \in E} p(o_{\tau'}^{t'}|w')^\alpha \sum\limits_{w'_p} \left(p(w'|w'_p)p(w'_p)\right)^\beta}$$

Dans cette équation, o_τ^t représente la séquence d'observations entre les instant τ et t, w_p un mot précédant w

et α, β sont des facteurs d'échelle entre les probabilités acoustiques et linguistiques.

2.2. Mesures locales

Pour les mesures locales nous introduisons des connaissances futures au mot analysé. Toutefois, ces connaissances se limitent à un voisinage local du mot et ne nécessitent pas la reconnaissance totale de la phrase prononcée. Un court délai est donc introduit afin d'attendre la disponibilité des informations nécessaires au calcul de la mesure. Par ailleurs, ces mesures sont fondées sur un autre critère probabiliste : l'estimation de la probabilité *a posteriori* d'un mot.
Le principe de cette mesure de confiance est de définir un voisinage autour du mot analysé $[w, \tau, t]$ en prenant en compte, de part et d'autre du mot, un nombre fixe de trames. Ainsi la taille totale en trames du voisinage V d'un mot w est la somme de la longueur du mot w et des longueurs des voisinages passé et futur. Les tailles des deux voisinages passé et futur sont indépendantes ce qui permet d'exploiter plus d'informations issues du voisinage passé, sans augmenter le délai introduit par le voisinage futur. Notons également que lorsque le voisinage futur est nul nous retrouvons une mesure trame-synchrone.
Une fois le nombre de trames du voisinage futur de V traité par le moteur de reconnaissance, nous extrayons du graphe de mots courant, engendré par le moteur de reconnaissance de manière trame-synchrone, le sous-graphe correspondant à V. Nous calculons sur celui-ci une estimation de la probabilité *a posteriori* du mot w, par la méthode *forward-backward* au niveau des mots résumée par les équations suivantes.
Soient $\Phi([w, \tau, t])$ et $\Psi([w, \tau, t])$ les probabilités *forward* et *backward* du mot $[w, \tau, t]$:

$$\Phi([w,\tau,t]) = p(o_\tau^t|w) \sum_{w_p} \sum_{\tau'} \Phi([w_p,\tau',\tau-1])\, p(w|w_p)$$

$$\Psi([w,\tau,t]) = p(o_\tau^t|w) \sum_{w_s} \sum_{t'} \Psi([w_s,t+1,t'])\, p(w_s|w)$$

w_p représente un mot précédant w et w_s un mot suivant w.
La probabilité *a posteriori* est alors estimée ainsi :

$$p(w|O) = p([w,\tau,t]|o_1^T) = \frac{\Phi([w,\tau,t])\Psi([w,\tau,t])}{\big(\sum_v \sum_\tau \Phi([v,\tau,T])\big) p(o_\tau^t|w)}$$

Dans le sous-graphe extrait, plusieurs occurrences du mot analysé peuvent apparaître à des positions temporelles similaires. La méthode *forward-backward* calcule donc la probabilité *a posteriori* de chacune des occurrences du mot analysé. N'en retenir qu'une seule sous-estimerait la vraie probabilité *a posteriori* du mot. Afin de gérer ce problème d'occurrences multiples, nous introduisons un facteur de flexibilité η et sommons les estimations des occurrences du mot analysé qui respectent certains critères dépendant de η.
Soient η le facteur de flexibilité, d la longueur du mot w et $[\widetilde{w}, \tilde{\tau}, \tilde{t}]$ une des occurrences de w appartenant au sous-graphe. Nous définissons les trois contraintes suivantes : $\tau - \eta\, d \leq \tilde{\tau} \leq \tau + \eta\, d$; $t - \eta\, d \leq \tilde{t} \leq t + \eta\, d$; $(1-\eta)\, d \leq \tilde{d} \leq (1+\eta)\, d$.
Soit F l'ensemble des occurrences d'un mot w respectant les contraintes précédentes, la confiance $C([w, \tau, t])$ de w est donnée par l'équation suivante :

$$C(w,\tau,t]) = \sum_{[\widetilde{w},\tilde{\tau},\tilde{t}]\in F} p([\widetilde{w},\tilde{\tau},\tilde{t}]|o_d^f)$$

o_d^f est la séquence d'observations correspondant au sous-graphe de mots associé à $[w, \tau, t]$ et à son voisinage V.

Remarque : Bien que les contraintes introduites pour les mesures locales et les mesures trames-synchrones semblent similaire, les rôles des facteurs ε et η sont différents : le facteur de relâchement sert à déterminer un ensemble d'occurrences de mots concurrents pour le rapport de vraisemblance, alors que le facteur de flexibilité permet de prendre en compte uniquement les occurrences du mot analysé dans le voisinage V , puis de sommer les valeurs de confiance de chacune d'elles.

3. CONDITIONS D'EXPÉRIMENTATION

3.1. Moteur de reconnaissance

Pour notre étude, nous avons choisi le moteur de reconnaissance grand vocabulaire *Julius* développé par des chercheurs de l'université de Kyoto [3]. Celui-ci présente plusieurs avantages : c'est un logiciel *open-source* offrant un très bon compromis temps-mémoire-précision tout en étant paramétrable ; de plus, il est compatible avec nos modèles acoustiques et linguistiques.
Lors du processus de reconnaissance, Julius construit un graphe d'exploration de manière trame synchrone, c'est à partir de ce graphe que nous estimons la valeur de confiance d'un mot.

3.2. Modèles acoustiques, linguistiques et lexique

Nous avons utilisé une paramétrisation acoustique du signal fondée sur les coefficients MFCC et une normalisation MCR (Mean Cepstral Removal).
Des modèles monophones basés sur des modèles de Markov cachés à trois états de type gauche-droit ont été appris à l'aide du logiciel HTK sur 40 heures d'émissions radiophoniques transcrites, extraites du corpus ESTER [2].
Le lexique et le modèle de langage ont été définis à partir d'un corpus composé de 16 années du journal français *Le Monde* et de la transcription manuelle de bulletins d'informations radiophoniques. Le lexique est constitué de près de 55 000 graphies différentes et le modèle de langage de 2,5M de bigrammes et 6M de trigrammes.

3.3. Corpus de développement et de test

Les corpus de développement et de test, d'environ une heure chacun, également extraits du corpus ESTER, sont indépendants des corpus d'apprentissages acoustiques et linguistiques. Les parties purement musicales et téléphoniques ont été supprimées. Chacun des corpus contient environ 11000 mots, avec un nombre moyen de mots par phrase sur le corpus de test de 11,5. Le taux d'erreur en mots de notre système de reconnaissance sur le corpus de test est de 33%.
Compte tenu de la taille des corpus de développement et de test, l'intervalle de confiance des résultats obtenus est de 0.8% avec un niveau de confiance de 95%.

4. RÉSULTATS

4.1. Méthodes d'évaluation

Nous avons évalué et comparé nos mesures de confiance à l'aide du taux d'égale erreur (EER). Pour cela, les mots reconnus par le système de reconnaissance sont étiquetés par notre mesure de confiance comme mots bons (*acceptés*) et mots faux (*rejetés*). Cet étiquetage dépend d'un seuil qui définit une frontière entre ces deux classes. Nous pouvons évaluer ainsi deux taux :

– le taux de *Fausse Acceptation* (FA) : le mot *accepté* était un mot incorrectement reconnu par le sytème de reconnaissance,
– le taux de *Faux Rejet* (FR) : le mot *rejeté* était mot correctement reconnu par le système de reconnaissance.

En faisant varier le seuil de décision, il est possible de déterminer une valeur particulière pour laquelle le taux de fausses acceptations et le taux de faux rejets sont identiques : le taux d'égale erreur EER.
Le seuil associé au taux d'EER a été mis au point sur le corpus de développement en fonction des paramètres ε, η, α et β. Puis ce seuil a été utilisé afin de calculer les taux de *Fausse Acceptation* et de *Faux Rejet* sur le corpus de test.

4.2. Mesure de référence

Dans cette étude nous comparons nos mesures de confiance entre elles mais également par rapport à une mesure de référence. Cette mesure est fondée sur une estimation de la probabilité *a posteriori* calculée sur l'intégralité du signal, le processus de reconnaissance ayant entièrement traité le document sonore. Cette mesure de référence est l'une des plus précises du domaine [7]. Nos mesures de confiance n'ayant qu'une connaissance partielle du signal traité, cette mesure de référence représente une limite à atteindre pour nos mesures. Sur le corpus de développement, la mesure de référence obtient un taux d'EER de 22,0%.

TAB. 1: Résultats obtenus par nos mesures de confiance et par la mesure de référence sur les corpus de développement et de test.

	corpus dév.	corpus test	
mesure	EER	FR	FA
référence	22,0%	21,2%	24,4%
locale début-60	23,2%	23,1%	23,2%
locale 84-84	23,0%	23,7%	24,5%
locale début-0	30,1%	30,3%	27,9%
trigramme	37,1%	34,5%	35,4%
trigramme inverse	37,1%	35,1%	37,4%
bigramme inverse	37,0%	35,1%	37,2%
bigramme	37,4%	36,6%	35,8%
unigramme	37,6%	38,8%	33,9%

4.3. Mesures trame-synchrones

Les premières évaluations sur le corpus de développement des mesures trame-synchrones ont montré que plus le degré du modèle de langage est élevé, meilleures sont les performances (EER de 37,6% pour le modèle unigramme à 37,1% pour le trigramme), ce qui s'explique par le fait que des informations linguistiques de plus grande portée sont prises en compte dans l'estimation de la mesure.

Nous avons également testé le remplacement d'un modèle de langage direct par un modèle de langage inverse (par exemple pour la séquence de mots $w_1w_2w_3$, $P(w_1|w_2w_3)$ remplace $P(w_3|w_1w_2)$ dans la mesure trigramme), tout en respectant la contrainte trame-synchrone de nos mesures. Ce remplacement apporte une amélioration dans le cas de la mesure bigramme mais pas dans celui de la mesure trigramme (cf. Tableau 1). Cette différence peut s'expliquer par le fait que la probabilité bigramme directe est utilisée par le moteur de reconnaissance et donc pour le choix des mots en compétition avec w. L'intégration du bigramme inverse dans le rapport de vraisemblance apporte une information supplémentaire. Le modèle trigramme n'étant pas utilisé par le moteur, l'apport du modèle trigramme inverse par rapport au modèle trigramme a donc peu d'effet.
Les modèles de langage bigramme et trigramme tiennent compte du mot précédent. Il est possible de choisir ces mots de différentes manières. Nous avons considéré deux méthodes : soit l'unique prédécesseur au sens de Viterbi, soit pour un mot $[w, \tau, t]$ l'ensemble de tous les précédents $[w_p, \tau_p, t_p]$ tels que $t_p = \tau - 1$ (méthode par défaut). Ne considérer que les prédécesseurs qu'au sens de Viterbi conduit à de moins bonnes performances (taux d'EER augmente de 3 à 4% en absolu).

Nous avons également testé d'autres variantes du calcul du rapport de vraisemblance. En effet, l'introduction du facteur de relâchement implique la prise en compte de multiples occurrences d'un mot w à des positions temporelles proches dans le graphe de mots. Prendre en compte les vraisemblances de toutes les occurrences ou seulement celle de score acoustique maximal donne le même taux d'EER.

4.4. Mesures locales

Nos mesures de confiance locales sont calculées sur un voisinage V d'un mot w défini par un voisinage passé et futur indépendant l'un de l'autre (cf. 2.2). Nous avons évalué l'influence de la taille de ces deux voisinages sur la précision de la mesure de confiance. La figure 1 présente le taux d'EER des mesures locales ayant un voisinage futur de 0, 40, 60 ou 84 trames (une trame=10ms) et un voisinage passé d'au minimum 40 trames et d'au maximum toutes les informations depuis le début du signal Nous pouvons remarquer que plus les voisinages sont de taille importante et plus la mesure obtenue est précise. Notamment, pour un voisinage futur fixe, le taux EER diminue fortement jusqu'à un voisinage passé de 84 trames. Lorsque le voisinage passé dépasse 84 trames, le taux diminue plus faiblement.
Mais surtout, nous pouvons remarquer que les mesures locales obtiennent des performances très proches de celle de la mesure de référence. En effet, avec seulement un voisinage de 84 trames de part et d'autre du mot le taux d'EER obtenu est de 23% contre 22% pour la mesure de référence (84 trames correspond à la durée moyenne de deux mots consécutifs dans le corpus de développement). Si l'intégralité du voisinage passé est pris en compte les deux taux d'EER sont identiques. Avec un délai réduit à seulement 60 trames (0,6s) et un voisinage passé complet, le taux obtenu est de 23,2%, très proche de celui obtenu par la mesure de référence.
Par ailleurs, si nous fixons une taille de voisinage futur nulle, la mesure locale obtenue est trame-synchrone. Le

résultat obtenu par cette mesure avec un voisinage passé commençant dès le début de la phrase est de 30,1%, valeur médiane entre notre meilleure mesure locale et nos autres mesures trame-synchrones. Cette différence s'explique par un nombre plus important de données prises en compte et par une méthode de calcul différente entre le rapport de vraisemblance (rapport entre séquences à nombre de mots identiques mais longueur différente) et la probabilité *a posteriori* (séquences de longueur identique mais nombre de mots différent).

4.5. Mesure de confiance vs. taux de mots corrects

Il est intéressant d'étudier la corrélation entre les valeurs prises par une mesure de confiance et la précision du système de reconnaissance. La moyenne des valeurs de confiance doit être proche du taux de mots correctement reconnus par le moteur sinon la mesure de confiance sous-estime ou sur-estime la fiabilité des mots.
Nous avons donc étudié la corrélation entre les valeurs de confiance calculées et le taux de mots corrects du système de reconnaissance, selon le principe suivant :
– estimation de la mesure de confiance de chacun des mots du corpus de développement,
– tri croissant des mots selon leur valeur de confiance,
– découpage uniforme de cet ensemble trié en N sous-ensembles de même taille,
– pour chacun de ces ensembles, calcul de la valeur de confiance moyenne et du taux de mots corrects obtenu par le système de reconnaissance.

La figure 2 illustre la répartition observée entre la valeur de confiance et le taux de mots corrects (TMC) sur le corpus de développement pour nos mesures trame-synchrone bigramme et locale 84-84 selon un découpage sur 20 ensembles. La courbe en trait continu et celle étoilée montrent la valeur de confiance moyenne dans chacun des ensembles, les deux autres le taux de mots corrects.
L'évolution croissante des deux courbes montre que nos mesures de confiance capturent bien une information de confiance mais cette corrélation est meilleure pour la mesure locale que pour la mesure bigramme. Nous avons exploité cette corrélation dans notre mesure de confiance bigramme en remplaçant la valeur de confiance calculée par le taux de mots correct de l'ensemble correspondant et avons noté une légère amélioration non significative du taux d'EER.

5. CONCLUSION

Les mesures de confiance que nous avons développées peuvent être calculées dès qu'une trame est traitée par le moteur de reconnaissance ou après un faible délai, permettant ainsi le traitement de flux audio continus sans avoir à attendre la fin de la reconnaissance. Nos mesures locales obtiennent des performances très proches de la mesure de référence avec un voisinage de seulement 84 trames de part et d'autre du mot (23% d'EER vs. 22%). Les mesures fondées sur la probabilité *a posteriori* sont plus précises que celles fondées sur le rapport de vraisemblance. De plus, le comportement de nos mesures est stable lors du passage du corpus de développement au corpus de test.
Les mesures que nous avons proposées peuvent être intégrées directement dans le moteur de reconnaissance ou bien utilisées par exemple pour améliorer la compréhension par des malentendants de transcriptions automatiques en direct.

RÉFÉRENCES

[1] S. Cox et S. Dasmahapatra. High-level approaches to confidence estimation in speech recognition. *IEEE Trans.*, pages 460–471, 2002.

[2] S. Galliano, E. Geoffrois, G. Gravier, J.F. Bonastre, D. Mostefa, et K. Choukri. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 315–320, 2006.

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

[4] J. Razik. *Mesures de confiance trame-synchrones et locales en reconnaissance automatique de la parole*. PhD thesis, Université Henri Poincaré, Nancy 1, 2007.

[5] C. Uhrik et W. Ward. Confidence metrics based on n-gram language model backoff behavior. In *EUROSPEECH*, pages 2771–2774, 1997.

[6] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, et A. Stolcke. Neural-network based measures of confidence for word recognition. In *ICASSP, Munich*, pages 887–890, 1997.

[7] F. Wessel, R. Schlüter, K. Macherey, et H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. SAP*, 9 :288–298, 2001.

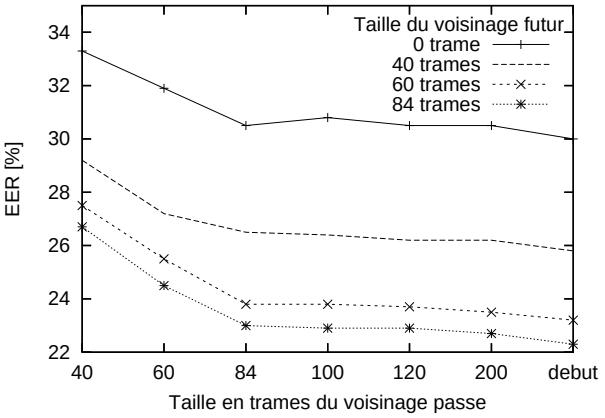

FIG. 1: Taux d'EER de la mesure de confiance locale selon les tailles des voisinages passé et futur.

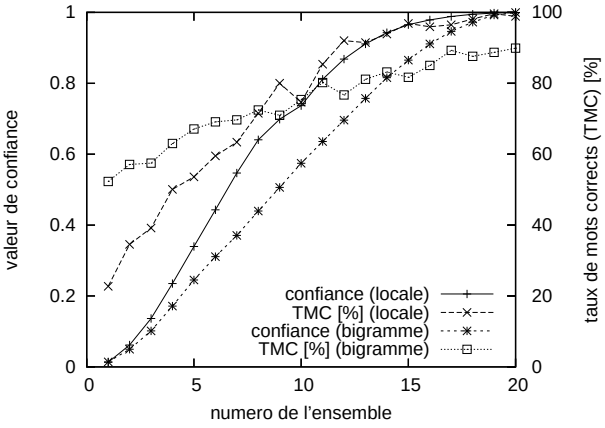

FIG. 2: Taux de mots corrects et valeur moyenne de confiance sur le corpus de développement pour les mesures bigramme et locale 84-84.