Transcription automatique pour malentendants : amélioration à l'aide de mesures de confiance locales

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

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

ABSTRACT

In this paper we present the use of confidence measures to improve the comprehension of automatic transcription by hard of hearing. The framework consists in live shows or live streams automatically transcribed by a large vocabulary speech recognition system. We have defined local confidence measures that can be estimated as soon as possible without having to wait for the recognition process to be completed. They have achieved results close to a reference post-processed measure computed on the whole signal and known to be the currently best accurate measure. We have then conducted an experiment to test the contribution of our confidence measure to improve the comprehension of an automatic transcription containing errors. We have thus introduced several modalities to highlight words of low confidence in this transcription and we have shown that these modalities can improve the comprehension of automatic transcription.

Keywords: speech recognition, hard of hearing, confidence measure, posterior probability

1. INTRODUCTION

Dans notre société moderne de plus en plus d'efforts sont faits pour aider les personnes handicapées à pouvoir vivre au quotidien comme tout un chacun. Toutefois ces aides concernent essentiellement les personnes à mobilité réduite alors que les autres formes de handicap comme la cécité et la surdité sont encore mal prises en compte. Ainsi la surdité, source d'exclusion sociale et familiale, devient un problème de plus en plus crucial. En effet, la France compte une population de plusieurs millions de sourds ou malentendants dont une part de plus en plus importante est constituée de personnes *devenues sourdes* (vieillesse, travail en milieu fortement bruité, amplification excessive de la musique,...). Les moyens de communications traditionnellement utilisés par les sourds sont des langages expressifs utilisant une gestuelle manuelle comme la langue des signes ou le langage LPC (Langage Parlé Complété). Pour les *devenus sourds* l'apprentissage de ces nouveaux langages est difficile et démotivant. Aussi leur préfèrent-ils le langage écrit.

Un effort a été demandé aux grands médias télévisuels pour fournir un sous-titrage de leurs émissions. Le sous-titrage à grande échelle comme le sous-titrage en direct nécessite l'utilisation d'un système de reconnaissance grand vocabulaire capable de produire une transcription automatique du flux audio. Cette reconnaissance automatique pourrait également apporter une aide importante aux élèves sourds et malentendants suivant des cours dans une salle de classe normale.

Néanmoins, les systèmes de reconnaissance automatique de parole (SRAP) ne sont pas infaillibles et les erreurs de reconnaissance peuvent engendrer des transcriptions difficilement compréhensibles par les malentendants. Nous nous sommes donc intéressés, d'une part, à la définition de mesures de confiance utilisables dans des applications en flux comme les deux précédents exemples, d'autre part, à l'amélioration que pourraient apporter de telles mesures dans la compréhension des transcriptions automatiques.

Cet article présente tout d'abord les mesures de confiance que nous avons définies, puis décrit le système de reconnaissance qui a permis de les mettre en oeuvre, avant de détailler une expérimentation concernant l'utilisation de l'un de nos indices de confiance pour améliorer la compréhension d'une transcription automatique.

2. MESURES DE CONFIANCE

L'objectif d'une mesure de confiance d'un mot est d'estimer la probabilité que ce mot ait été correctement reconnu par le système de reconnaissance de parole.
Nous avons défini deux types de mesures de confiance dites trame-synchrones et locales. Ces mesures permettent de calculer une valeur de confiance le plus tôt possible par rapport au processus de reconnaissance. Elles sont ainsi utiles dans le cadre d'applications en flux ou de diffusion en direct.

Les mesures de confiance trame-synchrones n'utilisent que les informations disponibles en même temps que 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.

Les mesures locales, elles, peuvent utiliser 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. C'est pourquoi ces mesures ne peuvent plus être qualifiées de trame-synchrones, mais, uniquement de locales.

Par ailleurs, ces deux types de mesures sont fondées sur deux méthodes d'estimation différentes : un rapport de vraisemblance pour les mesures trame-synchrones, la pro-

babilité *a posteriori* pour les mesures locales.

Nous avons évalué ces différentes mesures selon le critère du taux d'égale erreur (EER) sur une heure d'émissions radiophoniques issues du corpus ESTER [1]. Cette évaluation a montré que les mesures locales obtiennent de meilleures performances. Ainsi la mesure calculée sur un voisinage de 0,84 s de part et d'autre du mot considéré a obtenu des performances très proches d'une mesure très précise de l'état de l'art mais requérant la reconnaissance de la totalité de la phrase (23% d'EER versus 22%) [4]. Aussi avons nous décidé de ne tester que ces mesures locales dans le cadre de la transcription automatique pour malentendants. Nous allons donc maintenant détailler l'estimation de ces mesures.

Nos mesures locales sont donc fondées sur l'estimation de la probabilité *a posteriori* des mots.
Soit $[w, \tau, t]$ un mot commençant à l'instant τ et se terminant à l'instant t, le principe de la mesure locale 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. La figure 1 représente un tel voisinage V de w avec un voisinage passé de taille x et un voisinage futur de taille y.

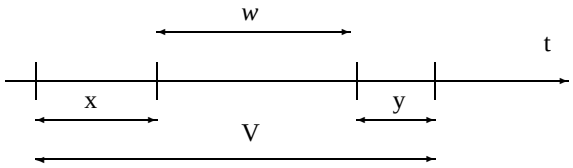

FIG. 1: Le voisinage V pris en compte pour le calcul de la mesure de confiance du mot w.

L'indépendance des tailles des deux voisinages de w permet de pouvoir exploiter plus d'informations issues du voisinage passé, sans augmenter le délai introduit par le voisinage futur.

Nous extrayons du graphe de mots, engendré par le moteur de reconnaissance, le sous-graphe correspondant à V et calculons sur celui-ci une estimation de la probabilité *a posteriori* du mot w, par la méthode *forward-backward* au niveau des mots détaillée dans [3] et résumée par les équations suivantes.

Soit $\Phi([w, \tau, t])$ et $\Psi([w, \tau, t])$ les probabilités *forward* et *backward* du mot $[w, \tau, t]$:

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

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

Dans ces équations, o_τ^t représente la séquence d'observation entre les instant τ et t, w_p un mot précédant w et w_s un mot suivant w. La probabilité *a posteriori* est alors estimée ainsi :

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

Sachant que la probabilité de l'observation $p(o_1^T)$ peut être calculée de la manière suivante :

$$p(o_1^T) = \sum_w \sum_\tau \Phi([w, \tau, T])$$

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.

Soit η le facteur de flexibilité, soit d la longueur du mot w et soit $[\widetilde{w}, \tilde{\tau}, \tilde{t}]$ 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.

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 [2]. Celui-ci présente plusieurs avantages : c'est un logiciel *open-source* offrant un très bon compromis temps-mémoire-précision tout en étant paramétrable [5] ; 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).

Les modèles de triphones basés sur des modèles de Markov cachés à trois états de type gauche-droit ont été appris à l'aide du logiciel HTK sur un corpus d'émissions radiophoniques transcrit d'environ 40 heures extrait du corpus ESTER. Les états des modèles triphones sont partagés à l'aide d'un arbre de décision.

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 60 000 graphies différentes et le modèle de langage de 19M de bigrammes et 28M de trigrammes.

4. TRANSCRIPTION POUR LES SOURDS ET MALENTENDANTS

Une alternative possible à la difficulté d'utiliser la lecture labiale, notamment dans le cas d'émissions audiovisuelles ou de cours, consiste en l'utilisation d'un système de transcription automatique. En effet, elle est utilisable par un grand nombre de personnes sans nécessiter l'apprentissage supplémentaire d'une nouvelle langue. La seule condition requise est de savoir lire, ce que savent faire la plupart des sourds ou malentendants, devenus ou de naissance. Cependant, un mot incorrect dans la transcription fournie par le moteur de reconnaissance peut rendre une phrase incompréhensible, perturber le lecteur et lui faire perdre sa concentration.

Nous proposons donc d'introduire nos mesures de confiance afin d'indiquer, par le biais de nouvelles modalités visuelles, la confiance à avoir dans les mots fournis par le système de reconnaissance.
Plus précisément, nous faisons l'hypothèse que le "marquage" des mots ayant un faible niveau de confiance par notre mesure permettra au lecteur de corriger plus facilement les erreurs issues du système de reconnaissance et d'améliorer la compréhension de l'ensemble de la transcription.
Nous avons donc réalisé une expérimentation dans laquelle les mots de la transcription ayant obtenu un faible niveau de confiance ont été transcrits :
– soit dans une autre couleur,
– soit dans un langage phonétique simplifié et dans une autre couleur.

4.1. Protocole expérimental

Dans cette expérience, nous souhaitions évaluer l'utilisation de mesures de confiance selon deux critères : l'amélioration de la compréhensibilité et l'appréciation de ces modalités de présentation. Le protocole expérimental que nous avons défini consiste donc à présenter à des sujets les transcriptions automatiques de quatre textes selon quatre modalités différentes et à faire exécuter aux sujets un test de ré-écriture et à leur poser des questions sur le contenu des textes.

Les quatre textes ont été conçus à l'origine pour connaître le niveau de lecture d'élèves de sixième. Les sujets abordés sont variés : conte, chronique sur les 24 heures du Mans, récit d'une expédition en avion, enquête policière sur le vol d'un ordinateur.
Chaque texte a été lu et enregistré par une même personne. Puis, ces enregistrements ont été transcrits par le système de reconnaissance décrit dans la section 3. Aucune adaptation au locuteur ou à l'environnement acoustique n'a été réalisée. Le taux de reconnaissance en mots sur l'ensemble des textes est de 71,4%.

Pour chaque texte transcrit, nous avons estimé la confiance de chaque mot reconnu à l'aide de la mesure de confiance locale utilisant un voisinage de 0,84s de part et d'autre du mot puisque cette mesure obtient de bonnes performances (cf. section 2).
La comparaison de ces valeurs de confiance à un seuil de décision, déterminé par le plus faible taux d'égale erreur (EER) obtenu sur un corpus de développement indépendant des quatre textes, permet de décider si un mot est correct ou incorrect (faible niveau de confiance) et ainsi de mettre ou non en évidence ce mot dans la transcription.

Pour chacun des textes, nous avons sélectionné une partie de la transcription automatique (240 mots en moyenne) que nous avons transcrite avec les quatre modalités suivantes :
– *brute* : la séquence de mots directement issue du système de reconnaissance sans aucune autre indication (encre noire) ,
– *oracle* : la séquence de mots avec les mots mal reconnus par le système de reconnaissance en couleur (bleu) ,
– *confiance* : la séquence de mots avec les mots jugés incorrects par notre mesure locale en couleur (bleu) ,
– *phonétique* : la séquence de mots avec les mots jugés incorrects par notre mesure locale transcrits phonétiquement en couleur (bleu) avec un alphabet phonétique simplifié. De plus, les mots incorrects contigus ont d'abord été regroupés puis transcrits phonétiquement afin de supprimer d'éventuelles mauvaises coupures issues du système de reconnaissance.

Le tableau 1 présente un extrait d'une transcription selon les quatre modalités ainsi que la séquence de mots initialement prononcée.

Les transcriptions automatiques ont été soumises à 20 sujets avec la tâche suivante à effectuer :
– deviner (reconstituer) le texte original d'une partie de la transcription correspondant à une soixantaine de mots,
– répondre à quelques questions de compréhension portant sur des mots plus ou moins bien reconnus par le moteur de reconnaissance,
– répondre à quelques questions subjectives sur l'appréciation de la modalité et l'estimation de la difficulté de la tâche.

Chacun des sujets a examiné les quatre textes mais avec une modalité différente par texte. Une durée limite de 15 minutes a été fixée par texte. Pour cette première expérimentation, le texte est présenté complet sans introduire d'effets de cadence dû au défilement de la transcription comme cela serait le cas pour un sous-titrage en direct. La mise en place d'un défilement synchronisé avec la voix sera expérimenté lors d'une autre phase de test.
Dans notre expérience, le test a été effectué par des sujets entendants car le concours de sujets sourds ou malentendants s'est avéré difficilement réalisable : nombre suffisant, disponibilité des sujets mais aussi des orthophonistes qui les accompagnent.

4.2. Résultats et discussion

Dans un premier temps, nous avons évalué le test de reconstitution du texte original en calculant le taux d'erreur en mots (taux d'insertions, omissions, substitutions) obtenu par chacun des sujets sans prendre en compte les fautes d'orthographe ou de grammaire afin de ne pas tenir compte du niveau de maîtrise du français des sujets. Le tableau 2 présente les résultats en taux d'erreur en mots pour chaque texte et chaque modalité. La colonne SRAP indique le taux d'erreur du système de reconnaissance automatique sur cette partie de texte.

Nous pouvons remarquer que sans aucune information complémentaire les sujets ont été capables de corriger une partie des erreurs faites par le SRAP. L'étude des autres résultats sera faite par rapport à ce taux d'erreur en mots obtenus par les sujets sur la transcription brute. Nous pou-

TAB. 1: Exemple d'une séquence de mots présentée selon les différentes modalités possibles.

Brute	Nous perdant de la hauteur une nourrice consigne continueront de percuter sommet
Oracle	Nous **perdant** de la hauteur **une nourrice consigne continueront** de percuter sommet
Confiance	Nous perdant de la hauteur **une nourrice consigne continueront** de percuter sommet
Phonétique	Nous perdant de la hauteur **u_n_n_ou_r_i_s_k_on_s_i_gn_e_k_on_t_i_n_u_r_on_t** de percuter sommet
Texte original	Nous perdons de la hauteur et nous risquons si nous continuons de percuter un sommet

TAB. 2: Taux d'erreur en mots sur les textes reconstitués par les sujets selon les différentes modalités.

texte	SRAP	brute	oracle	confiance	phonétique
Le Mans	18,8%	11,0%	9,0%	10,4%	**7,5%**
Conte suédois	31,3%	11,6%	16,4%	11,1%	**10,6%**
Première expédition	43,9%	40,4%	36,1%	**27,4%**	29,1%
Vol du PC	20,5%	17,4%	**12,3%**	16,4%	14,9%
Moyenne	29,0%	19,0%	18,2%	15,5%	**14,7%**

vons constater qu'en moyenne, le texte original est mieux retrouvé lorsqu'on donne une indication sur les mots faux (*oracle*) ou les mots susceptibles d'être faux au sens de la mesure de confiance (*confiance* et *phonétique*). Bien que le nombre de sujets par texte et par modalité (5) ne permette pas d'obtenir des résultats statistiquement significatifs, cela montre l'utilité de la mesure de confiance pour aider le lecteur à corriger une transcription erronée.
Il est également important de noter qu'en moyenne la modalité phonétique permet d'obtenir les meilleurs résultats. Deux raisons peuvent expliquer ces performances. Ecrire phonétiquement le mot de faible confiance plutôt que l'afficher simplement en couleur oriente moins le lecteur vers la recherche d'un mot de racine ou de sens proche de celui mis en couleur. Cette constatation a été faite également par les sujets eux-mêmes dans leurs avis sur l'utilisation de chaque modalité. La seconde raison est que le fait de concaténer les transcriptions phonétiques des mots de faible confiance contigus n'introduit pas de fausses coupures lexicales. Le lecteur peut ainsi retrouver plus facilement les mots originaux comme le montre l'exemple du tableau 1.

L'analyse des réponses aux questions subjectives d'appréciation montre que les sujets ont préféré la modalité *phonétique* celle-ci les aidant davantage à corriger les transcriptions. Il faut noter que la plupart des sujets n'avaient aucune connaissance en phonétique ni en traitement de la parole. Toutefois, le fait que les sujets aient préféré la modalité utilisant la phonétique doit être validé avec de vraies personnes sourdes ou malentendantes car les sujets entendants ou *devenus sourds* n'ont sans doute pas le même rapport à une mémoire phonétique que des personnes n'ayant jamais entendu.

Concernant les questions de compréhension portant sur des mots précis des transcriptions, une analyse approfondie montre que celles-ci ne sont pas pertinentes. En effet, pour la plupart des questions, la même réponse, juste ou fausse, a été donnée par l'ensemble des sujets.

5. CONCLUSION

Nous avons développé des mesures de confiance pouvant être calculées après un faible délai par rapport à la trame traitée par le moteur de reconnaissance, permettant ainsi le traitement de flux audio continus sans avoir à attendre la fin de la reconnaissance. De plus, ces mesures obtiennent des performances très proches d'une mesure très précise de l'état de l'art mais nécessitant la totalité de la phrase reconnue.
Nous avons également montré l'apport de notre mesure de confiance dans la compréhension d'une transcription automatique d'un flux audio puisque le fait de mettre en évidence dans cette transcription les mots de faible confiance a amélioré la compréhension de celle-ci.
De plus, écrire phonétiquement les mots susceptibles d'être faux donne de meilleurs résultats que le simple fait de les afficher dans une couleur différente. Toutefois cette conclusion reste à confirmer pour des sourds de naissance qui n'ont pas la mémoire phonétique des personnes ayant déjà entendu.
Par ailleurs, le seuil de décision utilisé pour déterminer les mots de faible confiance a été fixé à partir du taux EER, ne favorisant ainsi ni les fausses acceptations ni les faux rejets. Il n'est cependant pas établi que ce point de fonctionnement soit optimal pour ce type d'application. Une exploration d'autres points de fonctionnement fondés sur des critères perceptifs pourrait être menée afin d'évaluer l'influence de la proportion de fausses acceptations et de faux rejets sur la compréhension du lecteur.

RÉFÉRENCES

[1] 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.

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

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

[4] J. Razik, O. Mella, D. Fohr, et J.P. Haton. Frame-synchronous and local confidence measures for on-the-fly keyword spotting. In *ISSPA 2007*, 2007. 4 pages.

[5] T. Rotovnik, M.S. Maučec, B. Horvat, et Z. Kačič. A comparison of htk, isip and julius in slovenian large vocabulary continuous speech recognition. In *ICASSP*, pages 681–684, 2002.