Evaluation des Systèmes de Transcription enrichie

d'Emissions Radiophoniques (ESTER)

Plan d'Evaluation (phase 1)

Version 1.1

Dernière mise à jour le 21 novembre 2003.

1 Préambule

Ce document décrit le *plan d'évaluation de la phase 1* pour la campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques (ESTER).

Ce plan est une version modifiée du plan préliminaire d'évaluation (version 1.1) suite à la réunion du 24/04/2003 avec l'ensemble des acteurs du domaine. Le plan préliminaire, le compte-rendu de la réunion ainsi que des informations d'ordre général concernant la campagne sont disponibles sur le site *http://www.afcp-parole.org/ester*.

Le document présente tout d'abord très brièvement les objectifs scientifiques (de la première phase) de la campagne. Il décrit ensuite l'ensemble des tâches qui seront évaluées en divisant ces tâches en catégories. Les règles régissant la réalisation des différentes catégories de chacune des tâches sont énoncées dans cette partie du document. La description des tâches est suivie par une description des données de développement et de test et du calendrier de l'évaluation. Les règles d'engagement ainsi que les métriques et les formats d'échanges des données sont décrits dans des documents annexes disponibles sur le site de l'évaluation.

2 Objectifs

La campagne ESTER, organisée dans le cadre du projet EVALDA du programme Technolangue, a pour buts principaux de promouvoir une dynamique de l'évaluation en France, autour du traitement de la parole de langue française, de mettre en place une structure pérenne d'évaluation et de diffuser le plus largement possible les informations et les ressources concernées par ces évaluations.

L'axe prioritaire sera d'assurer un accès aux évaluations à un nombre aussi large que possible de participants.

Sur le plan scientifique, les résultats attendus sont bien évidemment de mesurer objectivement et de faire progresser les performances des systèmes de transcriptions enrichies en français, et d'inciter la fédération des efforts de recherche dans ce domaine.

L'objectif est également d'améliorer la visibilité du secteur de recherche concerné, par la mise en évidence du niveau de performance atteint par l'état de l'art, par la constitution d'un « club » d'acteurs identifiés et pouvant prouver leur niveau de compétence et par la publicité assurée au projet.

3 Tâches de l'évaluation

La phase 1 de l'évaluation s'articule autour des deux tâches :

- transcription orthographique (T)
- segmentation (S)

La tâche *extraction d'informations* ne sera envisagée que dans la phase 2 de la campagne. Chaque tâche est divisée en catégories [1], l'ensemble des catégories étant résumée dans le tableau 1.

tâche	catégorie	description
T	TRS	transcription orthographique
T	TTR	transcription temps réel
S	SES	suivi d'événements sonores
S	SRL	segmentation et regroupement de locuteurs
S	SVL	suivi de locuteurs

TAB. 1 – *Récapitulatif des tâches et catégories*

Pour chacune des tâches primaires, une catégorie est identifiée comme prioritaire (en gras dans le tableau 1) et les participants doivent impérativement retourner des résultats pour la catégorie prioritaire des tâches sur lesquels ils s'engagent.

3.1 Transcription

Cette tâche consiste à évaluer la transcription orthographique en sortie des systèmes, en terme de taux d'erreur de mots. Deux catégories de systèmes sont définies pour la tâche de transcription. Les participants s'engageant sur cette tâche doivent obligatoirement participer à la catégorie « transcription orthographique » (TRS) et sont encouragés à soumettre un maximum de résultats

1. Suite à la redéfinition des catégories, et afin d'éviter toutes confusions avec les catégories définies dans la version préliminaire du plan d'évaluation, la nomenclature des tâches et des catégories a été modifiée.

contrastifs. Un même système peut bien évidemment participer dans plusieurs catégories.

Pour l'ensemble des catégories, la normalisation (adaptation au locuteur, normalisation de scores, etc) ne peut se faire que sur la base d'un document. Un contraste utilisant une normalisation sur plusieurs documents est toujours possible mais ne pourra être présenté comme système primaire.

Les participants doivent fournir une description du ou des système(s) utilisés en spécifiant clairement les ressources linguistiques utilisées, les algorithmes et méthodes mis en œuvre ainsi que le temps CPU et la taille mémoire pour le décodage.

Transcription orthographique (TRS)

Afin de laisser le maximum de liberté aux participants, cette catégorie regroupe l'ensemble des systèmes de transcriptions orthographiques initialement divisés en plusieurs catégories selon les contraintes appliquées sur les ressources utilisées. Aucune restriction n'est donc a priori imposée aux systèmes pour la trancription. Les systèmes utilisant des ressources autres que celles fournies pour la campagne devront identifier ces ressources et fournir, dans la mesure du possible, des résultats contrastifs illustrant l'apport de ces ressources. Dans un souci de comparaison scientifique entre les différents constituants d'un système, les participants à la catégorie TRS sont encouragés à faire une évaluation en n'utilisant que les ressources distribuées pour la campagne [2]. Une évaluation des systèmes respectants ses conditions sera effectuées en plus de l'évaluation officielle. Les sites soumettant plusieurs contrastes devront identifier un système primaire sur la base duquel seront établis les résultats officiels.

Transcription temps réel (TTR)

Le temps de calcul pour le système complet est limité à 1 fois le temps réel sur un monoprocesseur standard. Cette catégorie vise à faire émerger des techniques alternatives et à favoriser des liens plus directs vers des démonstrateurs. Pour les site le désirant, une machine commune sera mise à disposition par la DGA (au minimum 2GHz de vitesse d'horloge et 1Go de mémoire vive). Un programme de calibrage de la puissance de la machine utilisée sera fourni afin de permettre une comparaison objective des temps de calcul.

3.1.1 Métriques

La métrique d'évaluation pour l'ensemble des tâches de transcription est le taux d'erreur de mot, calculé à partir des transcriptions de référence après normalisation. Les règles de normalisation des textes, les outils d'évaluation et les formats de soumissions seront détaillées dans un document annexe.

2. Pour des raisons historiques, on autorisera également pour ces systèmes l'utilisation du corpus BREF-120 pour la création de modèles acoustiques "bootstrap".

3.2 Segmentation

Cette tâche vise à évaluer les systèmes de suivi d'événements sonores ou de locuteurs ainsi que les systèmes d'indexation selon le locuteur. La tâche de segmentation se divise en quatre catégories, évaluées séparément. Les participants s'engageant sur cette tâche ainsi que sur la tâche de transcription doivent impérativement fournir des résultats concernant la catégorie "suivi d'événements sonores" (SES) et sont encouragés à participer à l'ensemble des catégories. Les participants ne s'engageant que sur la tâche de segmentation doivent obligatoirement participer à la catégorie "suivi et regroupement de locuteurs" (SRL).

Les participants doivent fournir une description du ou des système(s) utilisés en spécifiant clairement les ressources utilisées, les algorithmes et méthodes mis en œuvre ainsi que le temps CPU et la taille mémoire nécessaires.

Suivi d'événements sonores (SES)

Cette catégorie consiste à détecter les plages du flux audio pour lesquelles un événement sonore particulier est présent. Dans le cadre de l'évaluation, les deux événements étudiés seront la présence de parole et la présence de musique. Pour chaque événement, le système doit déterminer les plages contenant cet événement sur l'ensemble des documents du corpus de test [3]. Les données d'apprentissage pour cette catégorie sont limitée au corpus d'apprentissage de la phase 1.

Segmentation et regroupement de locuteurs (SRL)

Cette catégorie a pour but d'évaluer les algorithmes permettant de découper le flux audio en tours de parole et de regrouper les plages associées à un même locuteur (éventuellement non identifié). Le système retourne une segmentation du document spécifiant les plages de silence et un identifiant arbitraire de locuteur pour chaque plage contenant de la parole. Pour chaque document du corpus de test, la segmentation et le regroupement devront être établis sur la seule base de ce document. Dans cette catégorie, les participants doivent indiquer si le système utilise des connaissances a priori (par exemple, modèles appris a priori) ou pas.

Suivi de locuteurs (SVL)

Cette catégorie vise à évaluer les systèmes de suivi de locuteur qui permettent l'enrichissement de la transcription ou de la description du document. La tâche vise à détecter les plages correspond à un locuteur donné, connu à l'avance. Une liste des locuteurs à suivre est disponible sur le site Internet de l'évaluation. Comme pour SES, la tâche est une tâche de suivi. Pour chaque

3. Nous soulignons que cette tâche de suivi d'événement est légèrement différente de la tâche de classification classique qui consiste à segmenter en classes génériques silence, parole, musique, parole et musique. Dans cette évalution, il s'agit d'une détection parole/non-parole puis musique/non-musique.

locuteur de la liste, le système doit identifier les plages correspondant à ce locuteur dans l'ensemble des documents du corpus de test, un seul locuteur étant recherché à chaque fois. Pour chaque locuteur cible, les données d'apprentissage sont l'ensemble des occurences de ce locuteur dans le corpus d'apprentissage. Les locuteurs ne sont pas anonymes. L'utilisation de données autres que le corpus d'apprentissage, notamment pour le(s) modèle(s) de normalisation (modèle(s) du monde), est autorisée à condition d'être mentionné dans le descriptif du système. Comme pour la tâche de transcription, nous invitons les sites utilisant d'autres données à soumettre un contraste en n'utilisant que les données du corpus d'apprentissage.

Les performances seront moyennée sur l'ensemble des locuteurs à détecter.

3.2.1 Métriques

Pour les tâches de suivi d'événements (SES et SVL), la mesure de performance pour un événement donnée est le taux d'erreurs de détection, défini comme

$$\%\text{Err} = \frac{T(C|\overline{C}) + T(\overline{C}|C)}{T} \ ,$$

où $T(a|b)$ est le temps où l'événement a a été détecté comme b, C désigne l'événement recherché et $\overline{C}$ son absence, T étant le temps total.

Pour la tâche de segmentation et regroupement de locuteur, la métrique est le taux d'erreur définit comme la somme des taux de parole non détectée, de fausse détection de parole et mauvaise détection. Le taux de parole non détectée correspond aux portions de parole détectée comme silence. Inversement, le taux de fausse détection de parole correspond aux portions de silence pour lesquelles un locuteur a été détecté. Le taux de mauvaise détection correspond aux erreurs sur les identités (arbitraires) des locuteurs. Une correspondance entre noms de locuteurs et noms arbitraires fournies par le système est établie par appariement. Ces taux seront calculées sur l'ensemble des documents.

Pour l'ensemble des tâches de segmentation, les références seront établies à partir des transcriptions manuelles et d'une détection automatique des plages de silence. L'usage d'un détecteur automatique de silence permettra de marquer les plages de silence d'une durée supérieure à une seconde qui n'ont pas été marquées comme telles lors de la transcription manuelle. Le détecteur de silence utilisé sera fourni aux participants.

4 Ressources

L'ensemble des ressources utilisées pour la phase 1 de la campagne seront mises à la disposition des participants la DGA (ressources acoustiques) et par ELDA (ressources textuelles et lexicales). Avant d'avoir accès aux données, les participants devront signer un contrat d'utilisation des données avec ELDA (cf. document annexe sur les règles d'engagements et licence d'utilisation des données).

4.1 Ressources acoustiques

Les ressources acoustiques sont constituées d'émissions radiophoniques transcrites manuellement. Les émissions enregistrées sont des émissions d'information comportant le journal ainsi que des dossiers liés à l'actualité du moment. Pour la phase 1, les données sont issues de 2 sources différentes : France Inter et Radio France International (RFI). Le volume de donnée pour chacune de ces sources est spécifié dans le tableau 2.

source	train	dev	test
France Inter	19h40	2h40	2h40
RFI	11h	2h	2h
total	30h40	4h40	4h40
		40h	
période	1998–2000		

TAB. 2 – *Volume de données acoustiques et répartition train/dev/test.*

Les données France-Inter correspondent à la tranche matinale 7h–9h enregistrée du lundi au vendredi pendant 2 semaines en décembre 1998 (soit 20h) et au journal de 19h de Christophe Hondelatte, enregistré pendant 15 jours en juin 1999 (soit 5h). Les données RFI correspondent au tranches horaires 9h30–10h30 et 11h30–12h30, riches en accents variés, enregistrées en avril (3h), mai (8h) et septembre (4h) 2000.

L'ensemble des données, incluant les données du test à blanc, sera distribué dès le début de la campagne. Cependant, les participants sont encouragés à ne pas utiliser ni écouter les données identifiés comme test avant la date officielle de démarrage du test à blanc. Dans le cas contraire, il doit en être fait mention explicitement dans la description du système.

4.2 Ressources textuelles

Les ressources textuelles pour la phase 1 de la campagne correspondent aux années 1987 à 2002 du journal "Le Monde" augmentés du corpus MLCC contenant des transcriptions des débats du Conseil Européen.

4.3 Ressources lexicales

Il existe des phonétiseurs libres qui permettent la phonétisation du corpus d'apprentissage et du lexique[4]. En particulier, le Laboratoire Informatique d'Avignon met à la disposition des participants son phonétiseur, accessible depuis le site ESTER. De plus, la possibilité de mettre à la disposition des participants MHATLEX est à l'étude.

4. Cf. `http://tcts.fpms.ac.be/synthesis/mbrola.html`

5 Calendrier

28/05/2003	distribution des données
03/11/2003	début autorisé du test à blanc
17/11/2003	fin du test à blanc, envoie des résultats à la DGA
21/11/2003	validation des mesures de performances
15/01/2004	réunion de bilan

Les dates de début officiel du test à blanc et de retour des résultats sont susceptibles d'être modifiées.

6 Contacts

Pour plus de renseignements, contactez l'un des organisateurs:

- Guillaume Gravier (AFCP), `ggravier@irisa.fr`, 02 99 84 72 39
- Edouard Geoffrois (DGA), `Edouard.Geoffrois@etca.fr`, 01 42 31 96 68
- Kevin Mc Tait (ELDA), `mctait@elda.fr`, 01 43 13 33 33