ANTS : le système de transcription automatique du LORIA

Armelle Brun, Christophe Cerisara, Dominique Fohr,

Irina Illina, David Langlois, Odile Mella, Kamel Smaïli

Equipe Parole
LORIA
BP 239 54506 Vandœuvre-lès-Nancy, France
Tél.: ++33 (0)3.83.59.30.00 - Fax: ++33 (0)3.83.27.83.19
Mél: {brun,cerisara,fohr,illina,langlois,mella,smaili}@loria.fr - http://www.loria.fr/equipes/parole

ABSTRACT

In the context of the ESTER project, we have developed the ANTS system (Automatic News Transcription System). This paper presents the different modules : telephone-speech/broadband-speech segmentation, speech/music segmentation, breath detection, recognition engine. Preliminary results based on this system are put forward.

1. INTRODUCTION

La transcription automatique d'émissions radiophoniques pose des problèmes spécifiques. Citons les principaux : la longueur des fichiers de parole qui ne sont pas préalablement segmentés en phrases, les fréquents changements de locuteurs, parfois non-natifs, la superposition de parole et de musique, les alternances de parole large bande et de parole téléphonique, la présence de différents types de bruits et la parole simultanée.

L'équipe Parole du LORIA a développé un système de transcription automatique d'émissions radiophoniques : ANTS (*Automatic News Transcription System*) dans le cadre du projet ESTER [2]. Ce projet a pour objectif l'évaluation des systèmes automatiques de transcription de données radiophoniques francophones.

Nous allons tout d'abord décrire l'architecture générale du système avant de présenter le prototype réalisé dans le cadre de la campagne ESTER ainsi que quelques résultats d'expérimentation. Nous terminerons par les apports que nous souhaitons intégrer à cette première version.

2. LE SYSTÈME ANTS

Le système ANTS se compose de plusieurs parties : des modules de segmentation, un moteur de reconnaissance, des modèles acoustiques, un lexique et un modèle de langage (figure 1).

2.1 Les modules de segmentation

Le but de l'étape de segmentation du signal audio est double : d'une part, découper le signal audio en segments homogènes de taille acceptable par le moteur de reconnaissance, d'autre part, permettre d'utiliser des modèles ou des algorithmes spécifiques suivant la nature du segment.

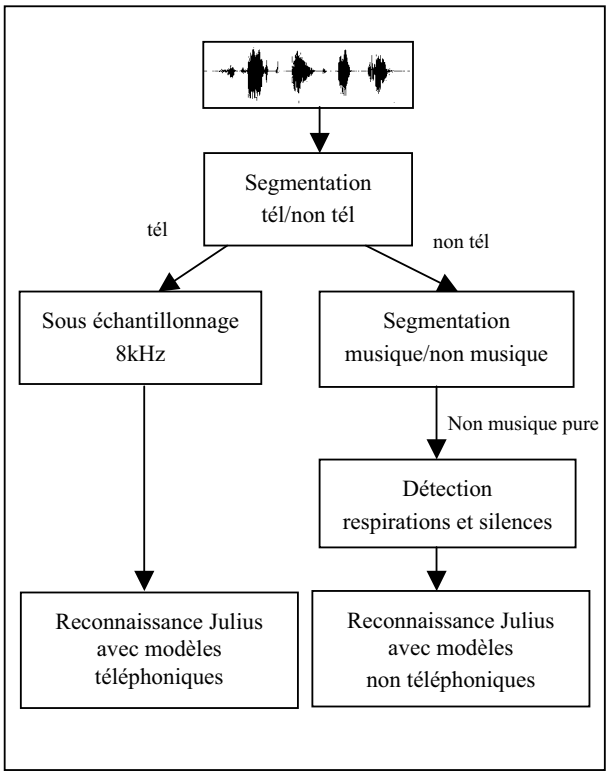

Figure 1 : Architecture du système ANTS

A la différence de Woodland [9], nous avons séparé la segmentation parole large bande / parole téléphonique de la détection des parties musicales. Ceci nous a permis de mettre en oeuvre une méthode spécifique pour la segmentation téléphone/non-téléphone. L'étape de segmentation est donc composée de trois modules qui s'enchaînent dans l'ordre suivant : la segmentation parole large bande / parole téléphonique puis la détection des parties musicales et enfin la recherche des respirations et des silences. A la fin de cette étape, nous obtenons des segments de taille raisonnable utilisables par le moteur de reconnaissance. Les paragraphes suivants décrivent les approches mises en œuvre dans ces trois modules.

La segmentation téléphone/non téléphone

Cette segmentation repose sur la différence d'énergie entre les basses [0 - 4kHz] et les hautes fréquences [4 kHz - 8kHz]. Cette différence doit être importante si le locuteur est au téléphone. Dans une première étape, le module affecte une valeur *e(t)* à chaque trame *t* de signal, *e(t)* vaut *1* si la différence d'énergie est supérieure à un seuil et *−1* sinon. Puis, la fonction *q(t)* est calculée à chaque trame *t* :

$$q(t)=\left|\sum_{t-1}^{t-L} e(t)-\sum_{t+1}^{t+L} e(t)\right| \qquad (1)$$

La courbe *q(t)*, plus ou moins lissée en fonction du facteur *L*, doit présenter un maximum lors du passage téléphone/non téléphone. Enfin, un algorithme de recherche de pics permet de trouver les points de rupture et de segmenter le flux audio.

La segmentation parole/ musique

Cette segmentation Parole/Musique est fondée sur la mise en compétition de cinq modèles constitués de mélanges de gaussiennes (GMM) : Parole Téléphonique (PT), Parole Non Téléphonique (PNT), Musique Instrumentale (MI), Chansons (C) et Parole-Musique (P&M), modélisant la superposition de la parole et de la musique. La segmentation est donc réalisée par une phase de reconnaissance dans laquelle une durée minimale de 0,5 seconde est imposée pour chaque segment [10].

La détection des respirations et des silences

Deux buts sont poursuivis à cette étape : premièrement, découper la parole en morceaux de plus petite taille ; deuxièmement trouver les groupes de souffle qui correspondent souvent à des entités syntaxiques ou sémantiques. Pour réaliser une telle segmentation, nous procédons à une reconnaissance au niveau phonétique en utilisant des modèles de phonèmes, un modèle de silence et un modèle de respiration ou de souffle. La grammaire utilisée pendant cette reconnaissance attribue la même probabilité à toutes les transitions entre les modèles.

Toute portion de signal comprise entre deux respirations et/ou silences est alors extraite.

2.2 Le moteur de reconnaissance, le lexique et le modèle de langage

Nous utilisons le moteur de reconnaissance Julius développé par Akinobu Lee [3]. Ce logiciel effectue la reconnaissance en deux passes : la première passe est trame-synchrone, elle utilise un bigramme et fournit un treillis (graphe) de mots. La deuxième passe est fondée sur un algorithme à pile et utilise un trigramme. Le système Julius offre la possibilité de définir plusieurs prononciations pour chaque mot du lexique.

3. MISE EN ŒUVRE DE ANTS DANS LA CAMPAGNE ESTER

Dans cette partie, nous allons décrire le système mis en place pour la première phase de la campagne ESTER et fournir les résultats de quelques expérimentations.

3.1 Données fournies pour la campagne d'évaluation

Pour cette campagne, des corpus acoustiques d'apprentissage, de développement et de test ont été distribués. Ils se composent d'émissions des radios France-Inter et RFI transcrites avec des transcriptions enrichies (texte, nom des locuteurs, description du fond musical…) réparties en 30 heures pour l'apprentissage, 4h40 pour le développement et 4h40 pour le test. Par ailleurs, 16 années du journal le Monde ont été fournies dans un format électronique pour l'apprentissage des modèles de langage [2].

Pour guider nos choix d'implémentation, nous avons utilisé plusieurs fichiers issus du corpus de développement fourni pour la campagne d'évaluation :

- 15 minutes d'un bulletin d'informations de France Inter que nous appellerons « fichier Inter »,
- une émission d'une heure de RFI que nous appellerons « fichier Rfi »

3.2 La segmentation téléphone/non téléphone

Dans un premier temps, le module de segmentation décrit au paragraphe 2.1 a été paramétré de la façon suivante : fenêtres de 16 ms décalées de 16 ms et le paramètre L fixé à 1s. Nous avons évalué ce module sur le fichier Rfi grâce à la segmentation manuelle réalisée par le CLIPS. Nous avons obtenu respectivement 99,8% de trames non téléphoniques et 99,9% de trames téléphoniques classées correctement, soit un total 0,20% de trames incorrectes. Ce résultat peut sembler satisfaisant mais, dans la pratique, les imprécisions dans la détection de début et de fin des parties téléphoniques peuvent provoquer une ou plusieurs erreurs de reconnaissance (mot coupé). Cette imprécision découle du lissage sur une seconde qui a été mis en place pour trouver des pics robustes. Nous avons donc mis en place une procédure pour affiner la détection de pics trouvés par l'algorithme initial. Pour cela, nous calculons, dans un intervalle de +/-0,5s autour du pic initialement trouvé, la différence d'énergie entre basse et haute fréquences avec un lissage de 0,3s. Puis nous recherchons dans cette courbe un maximum. Le test sur le même fichier donne 0,15% de trames incorrectes, les résultats sont détaillés Table 1. L'amélioration est nettement plus visible sur les taux d'erreur en mots obtenus par le module de reconnaissance. Ainsi, sur le fichier Inter, ce taux passe 31,9% à 30,3%.

Table 1 : Résultats de la détection téléphone /non-téléphone sur le fichier Rfi

	Total	non-téléphone reconnues téléphone	téléphone reconnues non-téléphone
Nb trames	225024	258	80
Nb secondes	3600	4	1

3.3 La segmentation parole/musique

Pour ce module, nous avons entraîné les 5 modèles PT, PNT, MI, C, et P&M, avec 16 gaussiennes en utilisant la boîte à outils HTK [1]. Ces modèles ont été appris sur la partie apprentissage du corpus ESTER sauf les modèles MI et C qui ont été appris sur des CDs audio. La paramétrisation est constituée de 36 coefficients : 12 MFCC, 12 Δ et 12 ΔΔ sans normalisation CMR (*Cepstral Mean Removal*).

Les segments reconnus comme musique instrumentale ou chanson sont définitivement éliminés.

3.4 La détection des respirations et des silences

Les modèles de phonèmes, de respiration et de silence ont été appris à l'aide de l'outil HTK sur la partie « non-téléphonique » de l'apprentissage du corpus ESTER. Ces modèles sont constitués de 3 états (gauche-droite, sans saut), avec 256 gaussiennes par état. La paramétrisation comporte 39 coefficients : 13 MFCC, 13 Δ et 13 ΔΔ et est complétée par une normalisation MCR réalisée en mode *off-line* sur chaque segment obtenu précédemment. Ces modèles et cette paramétrisation seront également utilisés par le moteur de reconnaissance.

Les respirations n'étant pas étiquetées dans le corpus d'apprentissage téléphonique, pour le moment, nous n'avons mis en place cette détection que pour la parole large bande.

3.5 Les modèles acoustiques et l'adaptation

Les modèles acoustiques

Nous avons créé deux ensembles de 36 modèles monophones, l'un pour la parole non-téléphonique, l'autre pour la parole téléphonique. Nous avons ajouté un modèle de respiration et un modèle de silence.

La paramétrisation acoustique, la topologie et l'apprentissage des modèles sont ceux présentés dans la section précédente.

Les modèles téléphoniques ont été appris à l'aide du corpus SpeechDat1000 [6] qui contient des phrases, des mots de commande et des nombres prononcés par 1000 locuteurs. La fréquence d'échantillonnage de ce corpus est de 8 kHz.

La Table 2 présente le taux de phonèmes correctement reconnus (*correct*) et le taux de reconnaissance phonétique (*accuracy*) obtenus sur des fichiers du corpus de développement d'ESTER avec la même probabilité de transition entre tous les phonèmes. Les modèles phonétiques sont testés sur la partie téléphonique des fichiers et les modèles non téléphoniques sur la partie non téléphonique. Nous pouvons observer une forte dégradation dans le cas de la parole téléphonique qui risque de se répercuter au niveau de la transcription automatique.

Table 2 : Résultats de la reconnaissance phonétique sur des données de développement

Modèles	Correct	Accuracy
non téléphoniques (16kHz)	75.0 %	70.4 %
téléphoniques (8kHz)	63.3 %	54.6 %

L'adaptation

Notre module de segmentation en locuteurs n'étant pas encore opérationnel, nous avons adapté ces modèles acoustiques de façon « aveugle ». Pour la reconnaissance du segment de parole courant, les modèles sont adaptés avec une méthode MLLR en utilisant les données acoustiques des segments précédent, courant et suivant. Rappelons que ces segments sont issus de la phase complète de segmentation. Plus précisément, une seule matrice de transformation, diagonale par blocs, a été utilisée pour l'adaptation MLLR.

Nous avons testé cette adaptation pour la transcription du fichier Inter. Nous avons obtenu 30,3% sans adaptation et 28,9% en adaptant les modèles.

3.6 Le lexique et le modèle de langage

Dans un premier temps, un lexique de 60000 mots a été construit à partir des mots les plus fréquents présents dans le journal « Le Monde » fourni dans le cadre de la campagne d'évaluation ESTER. La phonétisation des mots du lexique a été effectuée grâce à BDLEX [5] pour les noms communs et à un logiciel de phonétisation pour les noms propres qui ont été également vérifiés manuellement. A partir de ce lexique, nous avons calculé des modèles de langage bigramme et trigramme avec l'outil du CMU [4] en utilisant le Monde comme corpus d'apprentissage.

Pour simplifier, nous appellerons Modèle de Langage, le couple (lexique, modèles bigramme et trigramme).

Afin de minimiser le nombre de prononciations stockées dans le lexique, c'est-à-dire ne pas dupliquer le nombre de d'entrées, nous avons ajouté des modèles phonétiques élidables (*skippable*) comme le modèle de *schwa* élidable ou des phonèmes de liaisons élidables (/ R, n, z, t /).

Ce Modèle de Langage n'étant pas représentatif du style oral propre à la radio, nous en avons conçu un nouveau. Un nouveau lexique plus petit a été extrait à partir des années 1995-2001 du journal « Le Monde ».

Nous lui avons ajouté les mots apparus au moins trois fois dans les transcriptions des émissions radiophoniques du corpus d'apprentissage ESTER. Le résultat est un lexique de 55000 mots soit un total de 59000 prononciations, certains mots ayant plusieurs prononciations possibles. Le corpus d'apprentissage de ce nouveau Modèle de Langage a été construit à partir de la concaténation du corpus textuel du journal « Le Monde » 1995-2001 et de 10 fois le texte des transcriptions des émissions radiophoniques du corpus d'apprentissage ESTER.

L'utilisation de ce nouveau Modèle de Langage pour la transcription du fichier « Inter » permet d'obtenir un meilleur taux de reconnaissance : 31,9% de mots erronés contre 34,6% avec l'ancien modèle.

4. CONCLUSION ET PERSPECTIVES

Nous avons présenté une première version de notre système de transcription automatique ANTS. En effet, par manque de temps, nous n'avons pas pu ajuster tous les paramètres des modules développés ni intégrer tous les modules nécessaires dans un tel système.

En vue de la campagne d'évaluation ESTER qui aura lieu fin 2004, nous prévoyons d'intégrer de nouveaux modules, de tester de nouveaux paramètres et des modèles plus performants :

- Segmentation téléphone/non téléphone : nous souhaitons comparer notre approche avec une segmentation fondée sur des modèles GMMs ;
- Segmentation parole/musique : l'ajout de paramètres à long terme permettrait-il d'obtenir une meilleure segmentation ?
- Modèles acoustiques : nous envisageons de remplacer les modèles monophones par des modèles triphones Hommes/Femmes ;
- Adaptation : l'intégration d'un module de segmentation en locuteurs permettra de mettre en œuvre une adaptation plus fine (MAP, S-MLLR) ;
- Musique et bruits : actuellement, nous détectons les segments de parole sur fond musical mais nous n'utilisons pas cette connaissance. Il serait intéressant d'appliquer des méthodes de compensation ou des modèles acoustiques spécifiques ;
- Modèle de Langage : le modèle de langage utilisé actuellement est un langage élémentaire. Or, la parole radiophonique étant différente de celle que l'on peut avoir à partir d'un texte comme le journal « Le Monde », le nouveau corpus radiophonique nous permettra de mettre en place un modèle de langage plus adapté. Par ailleurs, pour la prochaine phase d'évaluation nous comptons développer un modèle de langage qui prend en compte le déphasage existant entre un début de segment et un début de phrase. En effet, certaines portions de phrase soumises à la reconnaissance ne respectent pas la structure langagière du français. Cela conduit à introduire dans le modèle de nouvelles structures qui n'ont pas été rencontrées en phase d'apprentissage. Nous comptons également utiliser des séquences de mots permettant de mieux reconnaître les suites de mots figées issues de la radio telle « Le Téléphone sonne » [7] [8].

REMERCIEMENTS

- Ce travail a pu être mené grâce au projet Technolangue EVALDA-ESTER.
- Nous remercions les équipes GEOD du CLIPS et TALNO du LIA pour la fourniture d'étiquetages manuels.

BIBLIOGRAPHIE

[1] S.J. Young and al., "*The HTK Book*", Cambridge, England ,Entropic Ltd., 1995.

[2] www.recherche.gouv.fr/technolangue

[3] A. Lee, T. Kawahara, and K. Shikano, "Julius – on open source real-time large vocabulary recognition engine", Eurospeech, pp. 1691-1694, 2001.

[4] P.R. Clarckson and R. Rosenfeld, "Statistical Language Modelling Using the CMU-Cambridge Toolkit", Eurospeech 97, pp. 2707-2710, Rhodes, Greece, 1997.

[5] M. De Calmès, G. Pérennou, « BDLEX : a Lexicon for Spoken and Written French », LREC 98, Grenade, pp. 1129-1136, 1998.

[6] French Speechdat(II) FDB-1000 (www.elra.info).

[7] I. Zitouni, K. Smaïli et J.P. Haton, « Statistical Language Modeling Based on Variable-length Sequences » Computer Speech and Language, vol.17 n°1, pages 27-41, 2003.

[8] A. Brun, K.Smaïli, et J.P. Haton, « Nouvelle approche de la sélection de vocabulaire pour la détection de thème »TALN, 2003.

[9] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk and S.J. Young, «Experiments in Broadcast News Transcription», ICASSP, 1998.

[10] J. Razik, D. Fohr, O. Mella, N. Parlangeau-Vallès, « Segmentation parole / musique pour la transcription automatique », JEP 2004.