Construction d'un corpus robuste de différents dialectes arabes

Mohamed BELGACEM

Laboratoire LIDILEM, Grenoble, France
Laboratoire UTIC, Tunisie
Tél. : ++33 (0) 6 33 88 18 98
Courriel : mohamed.belgacem@e.u-grenoble3.fr

ABSTRACT

This article is part of the project "Oréodule": a system-board real-time recognition, translation and speech synthesis Arabic. The object of our interest in this article is the presentation of a body of voice called Arabic. We detail the steps of establishing this body and the difficulties encountered during its development. We also integrated the practical results obtained during each phase (Record sizes, the total volume of our corpus, etc.).

1. INTRODUCTION

Notre article s'intègre dans le cadre du projet intitulé "Oréodule" : un système embarqué temps réel de reconnaissance, de traduction et de synthèse de la parole arabe. L'objet de notre intérêt dans cet article est la présentation d'un corpus vocal de la parole arabe. Nous détaillerons les étapes de constitution de ce corpus et les difficultés rencontrées lors de son élaboration. Nous intègrerons également les différents résultats pratiques obtenus lors de chaque phase (tailles des enregistrements, volume total du notre corpus, etc.).

2. PROBLÉMATIQUE

L'existence des dialectes de la langue constitue un défi pour le Traitement Automatique des Langues (TAL) en général, car il ajoute une autre série de variation de dimensions à partir d'une norme connue. Le problème est particulièrement intéressant, en arabe et ses différents dialectes. Toute approche réaliste et pratique du traitement de l'arabe doit rendre compte de l'usage dialectal, car il omniprésent. Pour mettre en évidence les différents phénomènes dialectaux pour la parole arabe et d'essayer de construire un système de reconnaissance automatique de dialecte arabe en utilisant les modèles GMM (Modèle de Mixture des Gaussiens). Pour aborder ce sujet il nous faut comme premier partie un corpus vocal. Or, le nombre limité des travaux dans ce domaine et l'inexistence de corpus vocal arabe commercialisé nous oblige à construire un tel corpus.

3. CONSTRUCTION DU CORPUS

3.1. Récupération et Enregistrement des Données vocales de la Parole Arabe

Concernant le recueil de données vocales en grande quantité pour la construction de notre corpus de la parole arabe avec ses différents dialectes, une approche intéressante consiste à « Télécharger » et « Enregistrer » un grand nombre d'émissions [Wai 04].

Les enregistrements de journaux radio- ou télédiffusés présentent un contenu varié : le signal acoustique peut correspondre à de la parole, de la musique ou du bruit, mais également à des mélanges de parole, de musique et de bruit. Ensuite il y a, pour la parole proprement dite, une grande diversité de locuteurs (Tunisien, Algérien, Marocain, Egyptien, Libanais, Syrien, Irakien, Yamin, Pays de Golf …) et de thèmes abordés (journaux, séries, débats politiques, sportif, éducation, social…). Plusieurs personnes peuvent intervenir sur un sujet donné successivement, voire simultanément. La qualité acoustique de l'enregistrement (fidélité) peut varier de manière considérable au cours du temps. La durée de tels enregistrements peut varier de quelques dizaines de secondes, minutes à plusieurs heures. Pour l'instant nous nous intéressons plus particulièrement aux nouvelles (journal, flash, revue de presse, incluant météo et bourse, économie, politique, faits de société ...) dans le document sonore. Toute autre forme d'enregistrement (publicités, jeux, fictions....) ne sera pas transcrite.

En suivant cette approche, nous avons enregistré l'équivalent de 10 heures de parole arabe de bonne qualité de différents dialectes à partir de 10 chaines TV et radios arabe.

Table 1: Statistiques de notre Corpus Vocal de la Parole Arabe

	Adulte	Enfant	Masc	Fém	Durée
Tunisien	100%	-	50%	50%	~ 90 mn
Algérien	90%	10%	55%	45%	~ 90 mn
Marocain	90%	10%	50%	50%	~ 90 mn
Egyptien	95%	5%	40%	60%	~ 92 mn
Palestinien	85%	15%	45%	55%	~ 60 mn
Libanais Syrien	95%	5%	50%	50%	~ 56 mn
Golfe *	100%	-	70%	30%	~ 160 mn
Somalien Soudanien	100%	-	100%	-	~ 21 mn
Non-Arabe*	-	-	-	-	~ 35 mn

Golf * **:** Ce groupe contient plusieurs pays (Irak, Koweiit, Arabie saoudite, Bahrein, Qatar, Yimen, Oman...).

Non- Arabe* : Anglais, Français, Iranien, Israélien…

3.2. Fitrages et Segmentation automatique selon les Locuteurs

Pour segmenter le signal de parole, il suffit de se placer à chaque position temporelle correspondant à un changement acoustique (changement de locuteur, silence, musique, parole, non parole…).

Pour aller plus vite et pour bien préciser toutes les informations de nos locuteurs. Nous avons utilisé un système de segmentation automatique en locuteurs Ce système est récupérer de l'équipe GETALP du laboratoire d'informatique de Grenoble [Vau 02]. Derrière cette segmentation automatique, on a fait une vérification manuelle pour améliorer les résultats et affectes à chaque locuteur les informations nécessaires (Nom, sexe, origine, dialecte, enregistrement en studio ou téléphonique, parole, non parole, music, publicité…).

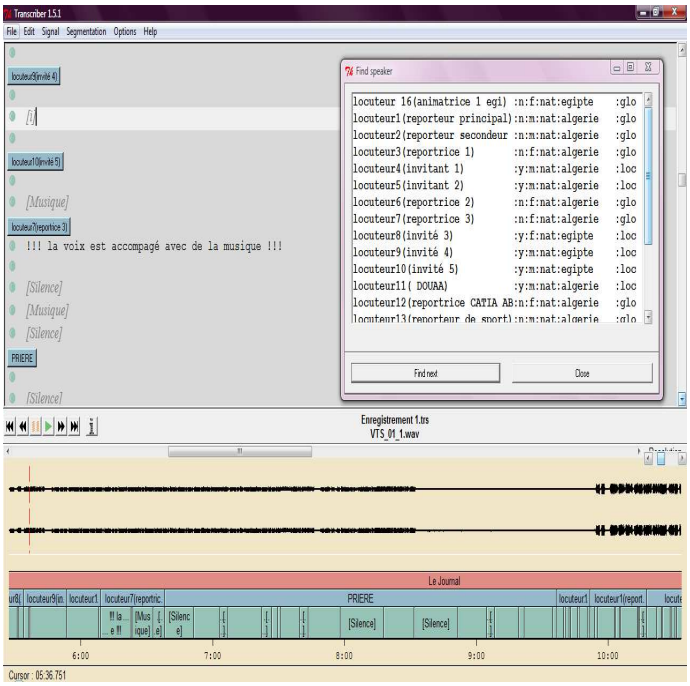

Figure 1 : Extrait d'une segmentation automatique et puis vérifier manuellement à l'aide de Transcriber.

3.2. Transcriptions de notre fichier wav à l'aide de Transcriber

Nous décrivons dans ce qui suit un ensemble de conventions pour structurer, annoter et transcrire des enregistrements de journaux radio- ou télédiffusés. Ces conventions doivent permettre de structurer les enregistrements au niveau du contenu thématique, des locuteurs et de la qualité acoustique. Les informations produites à ce sujet sont nommées annotations. La parole de chaque locuteur doit aussi être transcrite orthographiquement. C'est la transcription proprement

dite. La transcription est ici la partie la plus importante et donc sur laquelle le maximum d'attention doit être porté.

Les différentes étapes du travail de transcription sont : la segmentation de la bande son, l'identification des tours de paroles et des locuteurs, l'identification des sections thématiques, la transcription orthographique, et la vérification. Ces étapes peuvent être menées en parallèle ou au contraire appliquées séquentiellement sur de longues portions du signal, suivant le choix du transcripteur.

Dans notre cas on a réussit de faire que 37% de transcription de notre corpus (3heures et demie de transcription). Les émissions transcrites sont mélanges des dialectes arabes puisque ce sont des débats politiques.

Table 2: Pourcentage de transcription de chaque dialecte

Dialecte	Tu	AL	MA	EG	LIB SYR	YA	GO	IR
Trans %	5	6.5	7	4	5.5	3	5	1

Figure 2 : Extrait de la transcription à l'aide de Transcriber.

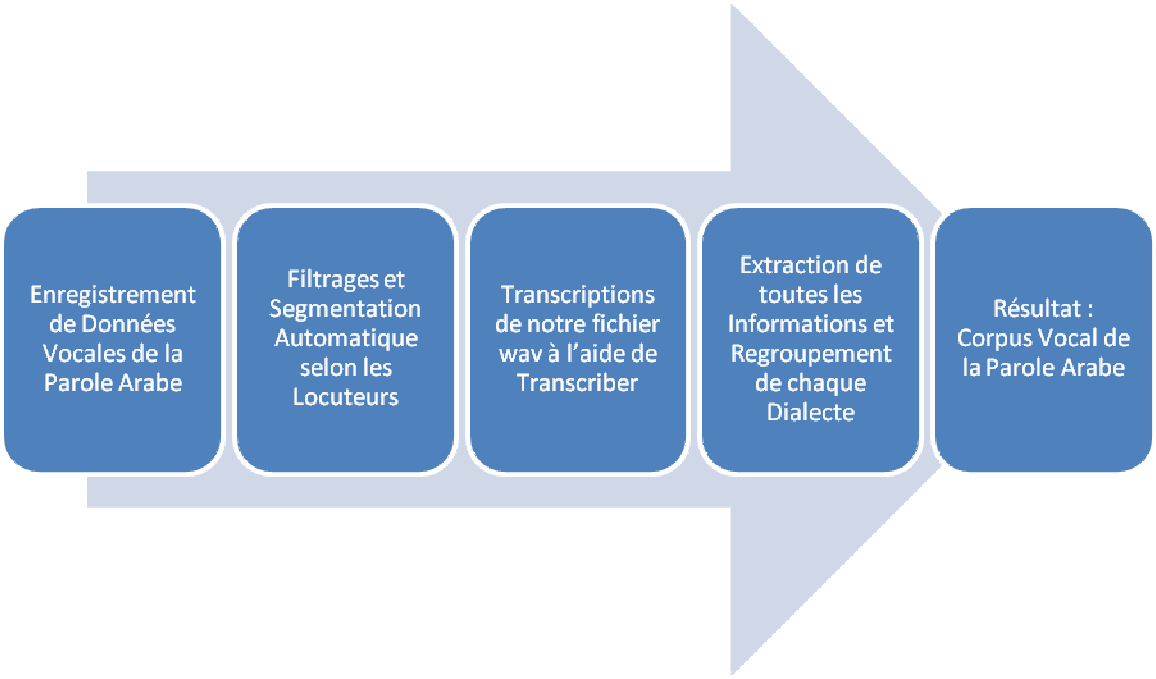

Figure 3 : Construction du corpus vocal pour la parole Arabe

5. CONCLUSION

Lors de l'élaboration du corpus, nous avons rencontré plusieurs difficultés. La majorité de ces contraintes est survenue lors de l'étape de segmentation et de transcription. Heureusement ces problèmes n'ont pas influé énormément sur les résultats de notre système de reconnaissance automatique de dialecte, lui-même basé sur ce corpus.

RÉFÉRENCES

[Vau 02] D. Vaufreydaz, Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue, Thèse de doctorat de l'Université J.Fourier - Grenoble I, France, 226 pages, Janvier 2002.

[Wai 04] A. Waibel, T. Schultz, S. Vogel, C. Fügen, M. Honal, M. Kolss, J. Reichert, S. Stüker, Towards Language Portability in Statistical Machine Translation, Special Session on Multilinguality in Speech Processing, ICASSP'04, Montreal, Canada, May 2004.

Sites Web arabophones :

http://www.aljazeera.net

http://www.tunisie.com/nouvelles