EMACOP : Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole

Dominique Vaufreydaz, Mohamad Akbar, Jean Caelen, Jean-François Serignat

EMACOP : Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole

Vaufreydaz Dominique, Akbar Mohammad, Caelen Jean, Serignat Jean-François

Laboratoire CLIPS (Communication Langagière et Interaction Personne-Système)
Université Joseph Fourrier, 38041 Grenoble cedex 9 France
Tél.: +33 (0)4 76 63 55 91
e-mail: [Dominique.Vaufreydaz, Mohammad.Akbar, Jean.Caelen, Jean-Francois.Serignat]@imag.fr

ABSTRACT

This paper presents an Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 95 and windows NT. It provides graphical interface functions to process the signal in order to compute acoustical parameters used in speech analysis, recognition and synthesis. Entries' extraction from main database is also provided from other sources. The new way of programming through the Microsoft COM (Component Object Model), which gives encapsulation/aggregation mechanisms and independence for each module, is used to increase reliability, flexibility and maintainability.

1. INTRODUCTION

La nécessité de disposer des bases de données sonores a émergé il y a une dizaine d'années, sous la poussée des méthodes de reconnaissance de la parole fondées sur l'apprentissage. Dans le milieu des années 80, des travaux ont été coordonnés en France ([Car84], [Des86]) autour de BDSONS, et prolongés dans des actions européennes (projets SAM I et SAM II). A l'heure actuelle, le besoin de données sonores reste encore une préoccupation essentielle.

Dans le même temps, les ordinateurs personnels (PCs) se sont répandus et ont atteint des performances intéressantes pour le traitement de la parole ; leur rapport prix/puissance n'a pas d'équivalent dans le monde des stations de travail Unix. Il est donc intéressant de les utiliser pour réaliser toutes les phases nécessaires à l'acquisition de signaux de parole et à leur stockage pour obtenir des environnements complets de traitement de la parole.

Dans le monde PC, de nombreux logiciels de portée générale existent et il aurait été plus rapide de coupler des utilitaires de bases de données et des logiciels de traitement de signal ([Akb97]) existants, pour construire un ensemble applicatif couvrant nos besoins. Mais l'inconvénient de cette solution réside dans le coût élevé de l'exploitation. En effet, ces logiciels sont souvent chers et leur utilisation, n'est pas, dans la plupart des cas, orientée vers un problème précis. En outre, cela aurait reporté une grande partie de la gestion (en particulier des fichiers) sur l'utilisateur. Enfin cela n'aurait pas permis de contrôler aisément et sans programmation les paramètres d'une session d'enregistrement. C'est pourquoi, malgré l'existence de logiciels comme EUROPEC [Zei91], développé sous DOS et utilisant une carte d'acquisition spécifique, nous avons développé une version plus étendue sous Windows 95 et Windows NT ne mettant en œuvre que du matériel standard : EMACOP (Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole).

La plupart du temps, les campagnes d'enregistrement mobilisent d'importantes ressources humaines pour guider ou assister les locuteurs dans leur tâche de diction, pour organiser l'enregistrement, pour préparer les scénarios et les données, etc. Il faut pouvoir contrôler les différents scénarios pour varier les conditions de capture : la lecture d'un texte ou d'une suite de mots ou de mots isolés, la répétition après écoute d'une phrase, le dialogue en réponse à des questions, etc. Les méthodes d'acquisition rigoureusement contrôlées sont donc lourdes. Elles nécessitent même un surcoût de travail, si l'acquisition numérique n 'est pas effectuée en direct, car il faut convertir et échantillonner les bandes analogiques et étiqueter les items sonores.

C'est pourquoi le développement d'un utilitaire de gestion et d'acquisition de grands corpus sur un matériel standard, nous semble d'un grand bénéfice à condition que la convivialité et la simplicité d'utilisation permettent à des non-spécialistes de maîtriser l'application.

2. L'APPLICATION

L'ensemble du logiciel se présente comme deux applications distinctes qui communiquent via un réseau grâce au protocole TCP/IP[1]. Elles sont toutes deux construites sur un modèle de composant (COM de Microsoft) qui les rend hautement modifiables et en facilite la maintenance : cela permet notamment l'accès à de multiples interfaces pour le même module, la recompilation d'un module sans retoucher aux autres, et il est possible de répartir de manière transparente

[1] Transfer Control Protocol / Internet Protocol

l'exécution des différents composants sur différentes machines.

La première application est le *serveur* qui permet la définition de la base de données et qui n'est exécutée que sur une seule machine. Elle fonctionne comme un serveur qui accepte les communications sur un "well known port" (qui peut être défini à chaque lancement) avec plusieurs clients.

La seconde application est le *client*. Elle a en charge toutes les fonctionnalités de présentation des items et d'acquisition du signal. Dans une optique d'exportation de savoir-faire du serveur vers le client, cela permet d'utiliser au maximum la machine locale et de décharger le serveur de tout traitement, autre que celui des données.

La communication entre le serveur et ses clients, se fait au moyen de requêtes, en mode connecté mais, dans le but de palier les problèmes réseaux, l'application cliente charge le maximum de données à la fois et peut, lors d'une coupure momentanée des communications, travailler avec les données qu'elle possède en local. Enfin si la durée de l'interruption dépasse un certain délai d'attente raisonnable pour l'utilisateur, il est possible d'envoyer plus tard au serveur le résultat de la session d'enregistrement en cours. Il sera alors possible de reprendre la session de travail là où l'on s'est arrêté auparavant. Un autre avantage de cette architecture semi-décentralisée est qu'elle peut supporter un nombre quelconque de clients dans la mesure où les transferts de données client-serveur peuvent être entrelacés avec l'enregistrement du signal sur le client.

Du fait qu'il existe deux applications avec des fonctionnalités différentes et que les niveaux de connaissance pour les gérer et les paramétrer sont différents, nous devons faire une distinction au niveau des compétences. Nous appelons **superviseur** la personne qui a en charge la gestion du serveur et de toutes les données qu'il contient, et **locuteur** toute personne qui utilise le logiciel client pour réaliser l'acquisition. Ces deux personnes peuvent évidemment être la même.

2.1. LE SERVEUR

La définition des données

La définition des données se fait à l'aide d'un utilitaire qui connaît plusieurs types d'entités : les corpus, leurs éléments, les locuteurs, les scénarios d'acquisition. Ces entités constituent les éléments de la base de données.

Après le lancement du serveur, une boîte de dialogue permet de choisir les modifications que l'on veut apporter à la base. L'interface se compose d'un enchaînement de fenêtres permettant la création des diverses entités et des liens entre celles-ci. Elle permet, non seulement l'entrée, mais aussi le contrôle de certaines données, qui sont, soit mises en forme automatiquement lors de la saisie, soit limitées, avant saisie, à un sous ensemble de valeurs (définitions standards) pour limiter les erreurs de frappe. Une meilleure cohérence du contenu de la base de données est ainsi garantie.

Le superviseur peut aussi, à tout moment, modifier les définitions standards en ce qui concerne les langues, les caractéristiques linguistiques et les types de corpus. Les définitions de base sont tirées de la norme ESPRIT-SAM [Tom91]. Toutes ces données sont de nature à pouvoir être utilisées dans le traitement automatique de la parole. Ainsi, pour un corpus, on définit son type (par ex. série de chiffres, de phrases, de textes, etc.), la langue et enfin la liste de ses éléments. Un locuteur est défini par son nom, son prénom, sa langue maternelle, sa caractéristique linguistique (par exemple parisien). Lors de l'exportation un code locuteur anonyme remplace le couple nom-prénom. A ceci s'ajoutent des caractéristiques physiques permettant de faire une différenciation plus précise des locuteurs : son âge (au moment de l'enregistrement), sa taille, son poids, son sexe et enfin s'il est fumeur ou non.

A l'aide de toutes ces informations, il est possible de rattacher divers renseignements aux signaux de parole enregistrés. La figure 1 schématise les données au sein de la base, les relations entre celles-ci et l'interface qui est présentée au locuteur pendant la tâche d'acquisition. Chaque locuteur connecté "voit" ainsi la partie de la base de données qui lui est affectée.

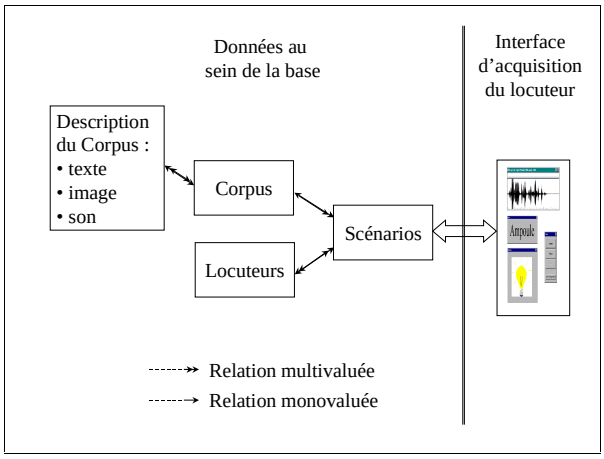

Figure 1 : Les données de la base gérées par le superviseur et l'interface du locuteur. Le superviseur définit les scénarios et gère la base de données (corpus, locuteurs et leurs descriptions).

Comme le montre la figure 1, la partie la plus importante du travail du superviseur est la définition des scénarios d'acquisition. Ces scénarios modélisent les plans de déroulement des sessions d'acquisition et permettent de définir les modalités d'affichage des stimuli. On notera que plusieurs scénarios peuvent être affectés à un locuteur, qu'à un scénario ne correspond qu'un seul corpus, lui même constitué d'un ensemble non limité de données.

La description d'un scénario est réalisée au moyen d'une interface qui propose au superviseur de faire des choix entre diverses alternatives et de définir plusieurs valeurs.

C’est ainsi qu’il détermine de quelle manière vont être présentés au locuteur les éléments à prononcer. Trois types de présentation sont disponibles. L’audition de ce qui doit être dit, l’affichage d'un texte ou l’apparition d’une image. Ces modes de présentation peuvent être utilisés de façon complémentaire. A ceci s’ajoute un temps de présentation qui détermine la durée de l’affichage des données en millisecondes. Enfin le superviseur indique le nombre de fois que doivent être répétés les items.

A ce niveau, il est possible de tenir compte des compétences du locuteur pour modifier le scénario :

- S'il ne possède aucune connaissance en matière de signal, on décide de lui faire énoncer les corpus sans qu’il n’ait aucune action sur ce qu’il produit,
- Il peut écouter ce qu’il a enregistré après chaque item, et juger par lui-même s'il le signal acquis est correct et s'il doit être archivé ou non,
- Il peut visualiser en temps réel ce qu’il enregistre pour en contrôler la pertinence,
- Il peut lui être alloué le droit de réaliser, pour chaque enregistrement, des modifications sur le signal correspondant à "couper" ou "modifier", pour éliminer les silences inutiles.

Le traitement des données

Le superviseur peut connaître, grâce à l'historique, toutes les actions exécutées par les locuteurs durant leurs sessions d’enregistrement. Cela lui permet de contrôler leur travail, s’il n’était pas présent ou si cela s’est fait *via* le réseau par exemple. Il trouve dans cet historique les annulations, les répétitions et les modifications que le locuteur a effectuées. Il a aussi accès à la connaissance du temps de réponse de ce dernier. Ce temps est calculé comme la différence entre l'instant d’affichage (ou de fin d’écoute) et l’instant de détection du début de parole.

Le superviseur peut faire des retouches de signal *a posteriori*. Il peut aussi vérifier ce qu’a fait et dit le locuteur. Ainsi les hésitations ou les substitutions de mots peuvent être traitées et/ou éliminées (ce système de correction a été utilisé pour la vérification de la base de données BREF [Lam91] par exemple). On possède ainsi à la fin de l'acquisition, une version des données épurées en vue de l’apprentissage ou de l’étiquetage.

Le superviseur peut exporter les signaux en vue de leur exploitation par d’autres logiciels. L'exportation se fait grâce à l’affichage de différentes arborescences (triées par locuteurs, par corpus ou par éléments) qui permettent aisément de choisir les sons à exporter selon les critères souhaités. Les formats SAM et ECHO [Kab96] sont par exemple supportés. Il est facile de définir son propre format d’exportation en créant un composant ayant la même interface que celui qui est livré avec l’application.

2.2. LES CLIENTS

Initialisation de l’application cliente

Après avoir défini entièrement toutes les modalités de présentation, un locuteur peut enregistrer. Il commence par s'identifier auprès du serveur dans le double but de lui attribuer une connexion et de lui délivrer les scénarios qui lui sont affectés. Il peut en choisir un, le mener à son terme ou l’arrêter en cours de route, le reprendre ou l'abandonner. Le locuteur peut sélectionner, dans un ordre quelconque, les scénarios d’acquisition qui lui sont demandés d'effectuer.

Un outil graphique de calibrage lui permet de déterminer les seuils de détection de sa voix par rapport à l’environnement sonore dans lequel il réalise son enregistrement. Cet outil lui permet de calculer automatiquement ces seuils et de les ajuster ensuite pour obtenir de meilleurs résultats. Ceci permet de ne pas être dépendant d’une interface avec des boutons pour déclencher l’enregistrement. On limite ainsi les effets de coupure et la présence de bruits, en début ou fin de signal, ce qui donne une meilleure qualité au signal produit même lorsque le locuteur n’a aucun droit de retouche sur celui-ci.

Interface générale pendant la phase d’acquisition

L’interface destinée au locuteur est présentée sur la fig. 2.

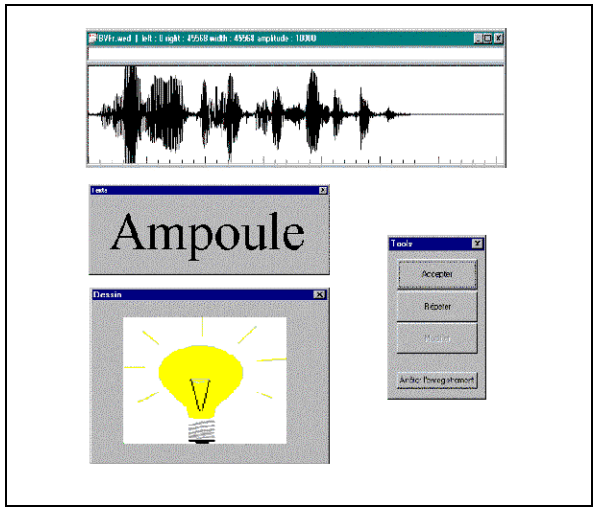

Figure 2 : Interface complète destinée au locuteur.

Elle est ici dans l'état où l'on présente un mot isolé *ampoule* et une image au locuteur. Dans cette situation, le locuteur a enregistré le mot et peut observer l'onde sonore acquise. Il peut éditer le signal et le modifier, puis il peut valider ou invalider l'enregistrement avec les boutons de la fenêtre de contrôle (ici l'enregistrement ne correspond manifestement pas à l'énoncé du mot *ampoule*).

Cette interface est composée de plusieurs fenêtres indépendantes, au maximum quatre selon les droits du locuteur (si le locuteur possède la visualisation et/ou

l'édition du signal produit et le mode de présentation par image et texte). De plus, il possède la visualisation du signal qu'il produit. La modularité de l'interface est totale puisque le locuteur peut placer à sa guise chaque fenêtre là où il le désire sur son écran.

La première qui se présente est celle contenant le menu de sélection. Elle lui permet dans un premier temps de lancer le scénario ou de revenir en arrière. Lors de la phase d'acquisition à proprement parler, un plus grand choix lui est proposé. Il peut accepter le signal qu'il vient de prononcer et passer au suivant. Il peut écouter ou voir le signal qu'il a produit (s'il en a le droit). Enfin il peut arrêter l'acquisition pour la reprendre plus tard.

La seconde fenêtre est celle qui présente le texte. Celui-ci est présenté dans une fonte de taille variable. L'affichage compte au maximum 4 lignes de 50 caractères (pour des raisons de lisibilité). Si l'élément dépasse cet affichage, des barres de défilement ("scrolling") apparaissent. Dans le cas où un temps de présentation est défini, ces barres sont remplacées par un défilement qui se déroule automatiquement et proportionnellement au texte.

La troisième fenêtre est celle qui autorise l'affichage des images. Celles-ci peuvent être de natures différentes (dessin ou image). Cela donne à EMACOP une portée d'utilisation en psycholinguistique par exemple : il peut été utilisé pour l'étude de normes associatives pour des images ou des sons, comme il en existe pour les mots [Mei98].

La dernière fenêtre est celle qui affiche, en temps réel, le signal pendant la phase d'acquisition et qui permet l'édition de celui-ci lorsque cela est possible. Elle propose un affichage sous forme d'onde (cf. figure 2) ou sous forme de nuage de points. La précision (nombre d'échantillons par pixel) peut être modifiée. Le locuteur peut, à l'aide d'un menu contextuel et d'une sélection à la souris, écouter tout ou partie du signal et réaliser des "couper" pour l'épurer.

2.3. VALIDATION

Le logiciel a été validé en situation de fonctionnement nominal sur une grande base de données. Les échanges à travers le réseau sont en cours de test.

3. CONCLUSION ET PERSPECTIVES

L'ensemble d'acquisition qui est décrit ci-dessus permet une très grande précision au niveau de la définition des différentes entités utilisées dans la base de données. Ainsi on peut être très rigoureux dans la définition des éléments et obtenir pour chaque scénario, une bonne adéquation entre le type de signaux désirés et la présentation des items. De plus, l'utilisation du réseau permet de faire des acquisitions en parallèle et donc facilite les grandes campagnes d'acquisition de corpus, éventuellement en parallèle sur différents sites. Enfin EMACOP autorise l'acquisition à grande distance (Internet) des corpus et permet enregistrer une plus grande diversité de personnes.

L'évaluation complète d'EMACOP est actuellement en cours et doit permettre de vérifier que l'on atteint un très bon contrôle de la prosodie, du débit et de la qualité de signal enregistré.

Enfin pour répondre aux attentes de nombreux utilisateurs, l'enregistrement en stéréo sera mis en place. Cela rendra possible l'enregistrement de dialogues, de fonds sonores mais aussi de signal provenant, par exemple d'un laryngophone.

BIBLIOGRAPHIE

[Akb97] Akbar M. (1997), « WaveEdit, An Interactive Speech Processing Environment for Microsoft Windows Platform ». EuroSpeech'97. Vol. 2, pp. 677-680.

[Car84] Carre R., Descout R., Mariani M., Rossi M. (1984), « The French Language Database : Defining, Planning and Recording a large Database ». Proc. IEEE ICASSP, San Diego, USA, 42.11.

[Des86] Descout R., Serignat J.F., Cervantes O., Carre R., (1986), « BDSONS : Une base de données des sons du français », 12th International Congress on Acoustic, Toronto, Canada, Vol. A, pp. 4-7.

[Kab96] Kabré, H. (1996), « ECHO: a Speech Input/Output system for the design of Interactive Speech-based Applications », American Voice Input/Output Systems, San-Jose, USA, pp. 179-188.

[Lam91] Lamel L.F., Gauvain J.L., Eskenazi M. (1991), « BREF, a Large Vocabulary Spoken Corpus For French », EuroSpeech'91.

[Mei98] Meiye J.P., Rouillard J., Vaufreydaz D. (1998), WebCompletion – Protocole de Normes Associatives sur Internet. Ecole Sémantique CNRS Caen 98. http://www-multicom.imag.fr/Cognition/Normes

[Tom91] Tomlison M.J. (1991), Guide to Database Generation – Recording Protocol, Final Version. SAM-RSRE-015, Marlvern, Angleterre.

[Zei91] Zeiliger J., Serignat J.F. (1991), Europec software (v4.1), User's Guide Release 4.1. SAM-ICP-045, Grenoble, France.