TALKAPILLAR : outil d'analyse de corpus oraux

Grégory Beller & Aurélien Marty

Ircam, Institut de Recherche et de Coordination Acoustique/Musique
1, place Igor Stravinsky
75004 Paris, France
tel : +33 (0) 1 44 78 48 75
Courriel : beller@ircam.fr & marty@ircam.fr

Abstract

In this paper we present a system devoted to speech for artistic purposes such as cinema, theater and contemporary music. It involves a relational database containing expressive and neutral French utterances. We describe the analysis system partly based on a concatenative Text-To-Speech system. A large set of descriptors of the segmented data permits a statistical approach of the speech phenomenon.

1. Introduction

Un système de synthèse musicale par concaténation d'unités, nommé CATERPILLAR [Sch03], a été élaboré au cours des dernières années. Le système a été étendu à un module « Text-To-Speech » (TTS) appelé TALKAPILLAR [Bel05]. L'une des applications de ce système est la reconstruction de la voix d'un locuteur cible, comme par exemple, une personnalité éminente disparue. TALKAPILLAR doit, dans ce cas, « prononcer » le texte exactement comme le locuteur cible l'aurait fait.

Le système comporte des outils d'analyse, de traitement et de synthèse de la parole, dans le but de répondre à des exigences artistiques. Des compositeurs de musique contemporaine s'intéressent à l'influence des émotions dans la voix et souhaitent explorer et utiliser de grandes bases de données. Un studio de doublage désire un système de synthèse de parole expressive. Enfin, certains metteurs en scène de théâtre souhaitent transformer, changer de type et synthétiser des voix sur scène.

La synthèse de parole par concaténation d'unités sélectionnées dans de grandes bases de données, appelée aussi « synthèse par corpus » [Hun96], est utilisée dans la plupart des systèmes TTS actuels pour la génération de la forme d'onde [pru01]. Depuis peu, le développement des méthodes par corpus et l'augmentation de la taille des bases de données ouvrent la voie, pour les systèmes TTS, à la synthèse de parole expressive. Certains auteurs [Bla03] ont enregistré plusieurs corpus de parole prononcée avec différentes expressivités et utilisent des méthodes TTS classiques séparément sur ces bases de données pour faire de la synthèse expressive. Une tentative de regroupement des différents corpus, sans séparation formelle, au sein d'une même base de données de parole a été réalisée [Bul02].

2. Vue d'ensemble du système

L'ensemble des processus mis en oeuvre dans l'analyse de la parole expressive est présenté schématiquement dans la figure 1. L'analyse de la parole expressive suppose de s'intéresser à deux informations véhiculées par la parole : la prosodie et la qualité vocale. Par exemple, la joie et la colère impliquent souvent une augmentation de la moyenne de la fréquence fondamentale. Cependant, elles diffèrent par leur qualité vocale : la colère présente un *jitter* plus fort. Ces deux niveaux d'information sont pris en compte dans TALKAPILLAR.

Le texte est d'abord analysé afin d'en extraire des informations d'ordre symbolique (transcription phonétique, nature grammaticale, etc.) [Mal98]. La séquence audio correspondante est segmentée par alignement et étiquetée selon l'expressivité.

Différentes analyses acoustiques sont déduites du signal, comme l'estimation de la fréquence fondamentale (f0). Des *valeurs caractéristiques* modélisant l'évolution temporelle de ces descripteurs acoustiques sont calculées à partir de la segmentation obtenue précédemment. Toutes ces données sont organisées, synchronisées et stockées dans une base de données. Enfin, une interface graphique efficace et conviviale permet l'exploration et l'analyse de la base, la synthèse par concaténation et la transformation basée sur le contenu de ces unités.

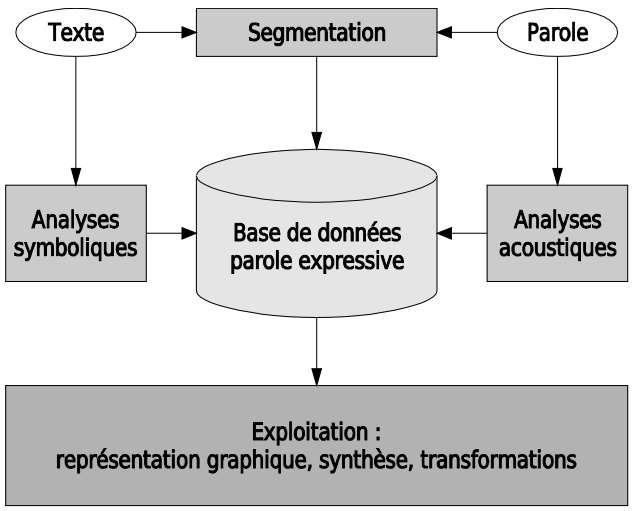

Figure 1 : Vue d'ensemble du système

3. Base de données

3.1. Contenu

Dans le cadre de nos travaux, nous avons progressivement mis en place une base de données contenant plus de six heures de parole en langue française, dont une partie est constituée d'un corpus de parole expressive prononcée par un acteur français. Ce corpus a été construit autour d'un jeu de 26 phrases de longueur variable, annotées (ponctuation, liaison, accentuation) dans le but de guider l'acteur et d'assurer une homogénéité structurelle de la phrase à travers ses diverses occurrences. En effet, Chacune de ces phrases a été prononcée avec les expressivité suivantes : *neutre, **colère, joie, tristesse, ennui, peur**, question neutre, dégoût, indignation, surprise positive et surprise négative*. Les expressivités « en gras » ont été simulées avec trois niveaux d'intensité par phrase : *faible, moyen, fort*. Une fois toutes les phrases enregistrées et traitées, l'acteur a librement écarté celles qui ne lui semblaient pas satisfaisantes. Au final, 539 phrases ont été retenues pour l'analyse.

3.2. Interface

L'utilisation de grandes quantités de données nécessite une architecture efficace proposant de multiples portes d'accès. Une base de données relationnelle DBMS, « *DataBase Management System »,* offre la possibilité de stocker facilement des fichiers audio, des centaines de milliers d'unités issues de leurs segmentations, leurs relations entre elles, et les méta-données les décrivant suivant plusieurs aspects. Cette base de données est accessible depuis Matlab via une interface graphique ou procédurale. Cette interface donne la possibilité de changer de DBMS ou d'utiliser d'autres bases de données de parole existantes. L'exemple ci-dessous montre la puissance et la simplicité d'emploi d'une telle interface :

» dbi('getunitdata', 'unit', dbs('getuidsfromsymbol', 'sOn', UnitTypes.syllable), FeatureTypes.f0, 'slope')

Cette commande retourne les pentes de évolution de la fréquence fondamentale de toutes les syllabes « sOn » présentes dans la base. On peut de même effectuer toute sorte de requête depuis Matlab, et au besoin, en filtrer le contenu. Par ailleurs, l'utilisateur peut explorer graphiquement le contenu de la base, accéder aux *valeurs caractéristiques,* écouter les différentes unités, etc. les figure 2, 3 et 4 ont été générées par quelques clics de souris. Signalons enfin que les données manipulées sont stockées dans le format SDIF (*Sound Description Interchange Format* [Sch00]).

3.3.Segmentation de la base de données

L'étape préliminaire à toutes les analyses acoustiques est la segmentation des séquences audio enregistrées en unités de taille variable. Une méthode d'alignement usuelle est utilisée. Elle consiste mettre en correspondance le texte et l'audio en plaçant des marqueurs temporels aux limites des différentes unités phonétiques. Une phrase de synthèse est bâtie à partir de diphones préalablement segmentés à la main, et correspondant à la transcription phonétique du texte. Les coefficients cepstraux calculés sur l'échelle de Mel MFCC des deux séquences audio (la séquence à segmenter, et la séquence synthétisée) sont alignés au moyen d'un algorithme DTW, « *Dynamic Time Warping »*. On obtient, la segmentation de la séquence en *semiphones, phones, diphones, syllabes, groupes prosodiques et phrases* (figure 2).

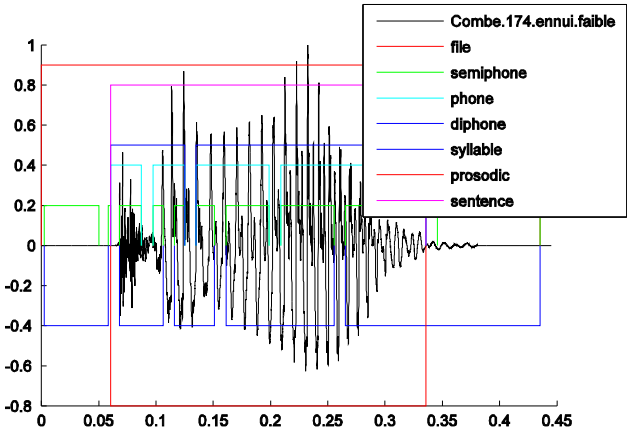

Figure 2 : Exemple de segmentation de la parole : la phrase « Comment ? ».

4. Descripteurs

Toutes les unités de la base sont étiquetées et renseignées au moyen de trois types de descripteurs :

- des descripteurs symboliques, catégoriels, décrivant le texte principalement ;

- des descripteurs dynamiques, liés aux évolutions temporelles de certains aspects du signal ;

- des descripteurs statiques, qui modélisent les descripteurs dynamiques pour chaque unité.

4.1. Descripteurs symboliques

Les descripteurs catégoriels rendent compte de l'appartenance d'une unité à une catégorie ou à une classe et permettent une hiérarchisation de la base (exemple : locuteur > homme > acteur, pour la hiérarchie « source sonore »). La description phonétique et syntaxique est obtenue grâce à EULER [Bag01] développé dans le cadre du projet MBROLA (TTS). Ce module réalise une analyse du texte incluant notamment la transcription graphème-phonème (traduit en XSAMPA), l'analyse grammaticale et la prédiction des liaisons et des syllabes accentuées. Cette dernière analyse permet de créer des frontières prosodiques : un groupe prosodique étant défini arbitrairement par une séquence de syllabes non accentuées suivie d'une syllabe accentuée [Mal98]. Une étude sur de la parole neutre a montré qu'une syllabe accentuée est en moyenne deux fois plus longue qu'une syllabe non accentuée. Beaucoup d'autres descripteurs

sont déduits de la place relative de l'unité dans la phrase, dans le mot et dans la syllabe.

4.2. Descripteurs dynamiques

Les descripteurs dynamiques rendent compte de l'évolution temporelle d'un aspect du signal, comme par exemple sa fréquence fondamentale.

4.2.1. Débit: Le débit de parole est souvent déni comme le nombre moyen de syllabes par seconde, calculé sur une phrase ou plus. D'après nos observations, l'accentuation est liée à de fortes variations de la durée des syllabes. C'est pourquoi nous préférons considérer le débit de parole comme un phénomène local, représenté par la séquence des durées des syllabes. Ainsi, la courbe dynamique du débit est une interpolation linéaire de cette séquence de durées syllabiques (figure 3). Une décélération correspond à une montée de la courbe et une accélération à une descente.

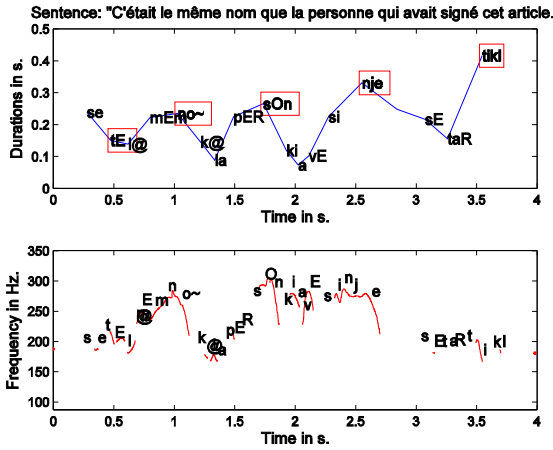

Figure 3 : Durée des syllabes et fréquence fondamentale de la phrase : « C'était le même nom que la personne qui avait signé cet article. », exprimée avec joie.

La figure 3 montre les informations suivantes :

- les syllabes encadrées sont considérées accentuées par EULER ;

- la courbe de débit présente des maxima locaux correspondant à des décélérations et une pente globale positive sur l'ensemble de la phrase ;

- l'accent final parait plus facilement identifiable sur la courbe de débit que sur la courbe de f0.

Notons que dans cet exemple, la prédiction des syllabes accentuées concorde à la performance de l'acteur bien qu'EULER ait été destiné à la parole neutre. C'est parce que celui-ci a tenté de reproduire les frontières prosodiques annotées selon EULER en amont. Le principal avantage de cette définition du débit est son aspect dynamique pouvant être relié au rythme.

4.2.2. Fréquence fondamentale et énergie: L'estimation de la fréquence fondamentale est réalisée par l'algorithme YIN [Che02]. Ce dernier donne également l'énergie et le rapport harmonique sur bruit (indice de voicement) du signal sur chaque trame. En seuillant judicieusement la courbe de voisement, on interpole sur les segments non voisés les estimations de f0 sur les parties voisées.

4.3. Descripteurs statiques

Les descripteurs statiques servent principalement à modéliser l'évolution temporelle des descripteurs dynamiques. Un vecteur de valeurs caractéristiques est représenté à la figure 4. Il se compose de :

- la moyenne arithmétique, géométrique, est l'écart-type ;

- le minimum, le maximum, l'écart absolu, la pente, qui donne l'évolution « brute » du descripteur, et le rayon de courbure (approximation polynomiale de Legendre du 2e ordre) ;

- les valeurs et pentes au début et à la fin de l'unité (pour évaluer des coûts de concaténation)

- les centres temporels de gravité et d'anti-gravité donnant l'instant de la plus importante élévation ou dépression de la courbe, ainsi que ses 4 premiers moments temporels

- le spectre de Fourier du descripteur normalisé dans 5 bandes et ses 4 premiers moments. Cela révèle si le descripteur possède des variations rapides ou lentes, ou s'il oscille (mesure du *jitter* et du *shimmer*).

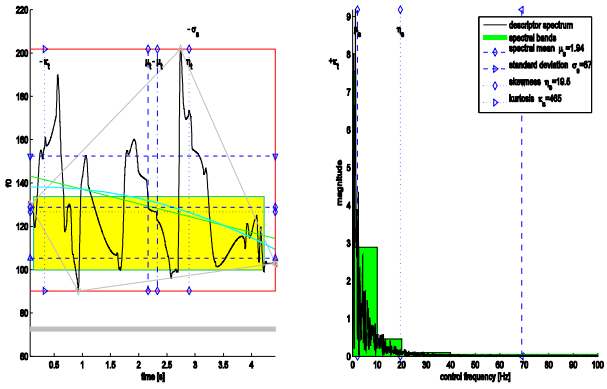

Figure 4 : Exemple de valeurs caractéristiques de la fréquence fondamentale calculées sur une phrase exprimée avec colère.

5. Exploitation des bases de données

5.1. Synthèse

Le système de synthèse TALKAPILLAR est principalement destiné, contrairement à la plupart des TTS classiques, à (re)produire une expressivité spécifique. L'utilisateur peut agir sur chaque étape du processus de création afin d'avoir un contrôle total du résultat final, essentiel pour la création artistique. Une interface graphique permet de paramétrer la synthèse et d'appliquer des transformations au signal synthétisé. La sélection des unités repose sur leurs descriptions. L'adéquation des unités au contexte désiré (coût de cible) et la qualité de la séquence formée (coût de

concaténation) sont assurées par la minimisation d'une combinaison linéaire de distances sur les descripteurs. La meilleure séquence est donnée par un algorithme de Viterbi.

5.2. Synthèse hybride

L'architecture commune du synthétiseur TTS TALKAPILLAR et du synthétiseur de phrases musicales CATERPILLAR [Sch03], permet de réaliser des synthèses hybrides mêlant parole et musique. La flexibilité du paramétrage de la sélection (descripteurs utilisés, poids associés) confère à l'utilisateur une grande liberté de création. Par exemple, des phrases musicales ont été synthétiser à partir d'une prosodie verbale cible [Bel04].

5.3. Transformation

Différentes transformations peuvent être appliquées aux unités sélectionnées, dépendantes du contexte La transformation basée sur le contenu permet de modifier localement des paramètres en fonction de l'unité à transformer et de son contexte. Certaines transformations (alignement de période, fondue croisée) ont pour but d'améliorer la synthèse en lissant la concaténation. Les dilatation temporelle et transposition du Super Vocoder de Phase [Boo04] permettent de modifier la prosodie et l'expressivité de la voix.

5.4. Analyse et exploration graphique

La quantité d'information disponible peut-être à tout moment consultée via une interface graphique. Des analyses statistiques nécessitant de nombreux individus sont facilitées par l'utilisation de Matlab. Différents résultats concernant l'influence des expressivités sur le débit (syllabique, pauses, respirations) et f0 sont présentés dans [Bel06].

6. Conclusion

À travers cet article, nous avons présenté un système conçu pour analyser, transformer et synthétiser de la parole expressive. TALKAPILLAR est un outil efficace permettant, via une interface simple, l'exploration statistique de bases de données contenant une grande quantité d'informations. Nous l'utilisons dans le cadre d'applications artistiques en rapport avec le cinéma, le théâtre et la musique contemporaine. Des exemples sonores sont disponibles à l'adresse suivante : http://recherche.ircam.fr/equipes/analyse-synthese/concat.

Nos travaux futurs se concentrent à présent sur la qualité de la synthèse par l'amélioration de toutes les étapes citées précédemment. Cela requiert notamment la possibilité de décrire et de modifier la qualité vocale.

7. Remerciements

Les auteurs tiennent à remercier l'acteur, Jacques Combe, pour sa performance.

Bibliographie

[Bag01] Bagein M., Dutoit T., Tounsi N., Malfrère F., Ruelle A., & Wynsberghe D. (2001), *Le projet EULER, Vers une synthèse de parole générique et multilingue*, Traitement automatique des langues, vol. 42, no. 1.

[Bel04] Beller G. (2004), *La musicalité de la voix parlée*, maîtrise de musique, Université Paris 8, Paris.

[Bel05] Beller G., Schwarz D., Hueber T. & Rodet X. (2005), *Hybrid concatenative synthesis in the intersection of speech and music*, JIM, vol. 12, pp. 41-45.

[Bel06] Beller G., Schwarz D., Hueber T. & Rodet X. (2006), *Speech Rates in French Expressive Speech*, Speech Prosody 2006, Dresden.

[Bla03] Black A. (2003), *Unit selection and emotional speech*, Eurospeech.

[Boo04] Bogaards N., Roebel A. & Rodet X. (2004), *Sound analysis and processing with audiosculpt 2*, ICMC, Miami, USA.

[Bul02] Bulut M., Shrikanth S., Narayanan S., & Syrdal A. K (2002), *Expressive speech synthesis using a concatenative synthesizer*, ICSLP, New Jersey.

[Che02] de Cheveigné A. & Kawahara H., (2002), *YIN, a Fundamental Frequency Estimator for Speech and Music*, JASA, vol. 111, pp. 1917-1930.

[Hun96] Hunt A. J. & Black A.W. (1996), *Unit selection in a concatenative speech synthesis system using a large speech database*, ICASSP, Atlanta, GA, pp. 373-376.

[Mal98] Malfrère F., Dutoit T., & Mertens P. (1998), *Automatic prosody generation using supra segmental unit selection*, SSW3, pp. 323-328.

[Pru01] Prudon R. & d'Alessandro C. (2001), *A selection/concatenation TTS synthesis system : Databases developement, system design, comparative evaluation*, Speech Synthesis Workshop, Scotland.

[Sch00] Schwarz D. & Wright M. (2000), *Extensions and Applications of the SDIF Sound Description Interchange Format*, ICMC, Berlin, pp. 481-484.

[Sch03] Schwarz D. (2003). *New Developments in Data-Driven Concatenative Sound Synthesis*, ICMC, Singapore, pp. 443-446.