Master 2 SAR-ATIAM, IRCAM
Université Pierre et Marie Curie, Paris VI
Magistère EEA, ENS de CACHAN

Étude et modèle génératif de l'expressivité dans la parole

Rapport de stage

Beller Grégory Lundi 27 Juin 2005

Plan:

Résumé:

La capacité d'exprimer et d'identifier des émotions, des intentions ou des attitudes par la modulation de caractéristiques de la voix est fondamentale dans la communication humaine. Elle coordonne, en particulier, les interactions sociales avec les bébés, ainsi que des jeux de langue (donnant la rétroaction, réclamant l'attention). Il semble bien que tous ces aspects maîtrisés ou non de la prononciation d'une phrase recouvrent plus d'une catégorie. Pour désigner cet ensemble, nous utiliserons dans la suite le terme d' « expressivité » tout en sachant qu'il faudra bien distinguer ces catégories. Les émotions par exemple ont quelques effets mécaniques sur la physiologie, comme la modulation de la fréquence cardiaque ou la sécheresse dans la bouche, qui ont à leur tour des effets sur l'intonation de la voix. Ainsi il est, en principe, possible d'extraire l'information émotive d'une phrase à partir de ses caractéristiques acoustiques, dont sa prosodie.

Dans le domaine artistique, de nombreux compositeurs (Emmanuel Nunes, Jonathan Harvey, Alain Bonardi…) et metteurs en scène (Jean-François Perret…) s'intéressent aujourd'hui aux multiples possibilités que pourrait fournir un système d'analyse, de transformation et de synthèse de l'expressivité dans la voix parlée. C'est le but que ce stage se propose d'atteindre. C'est aussi le but de la thèse qui va le poursuivre.

I. Présentation:

I.1. Émotions Analysées:

Selon Picard (Picard, 1997) les émotions primaires concernées sont définies par des catégories discrètes (approche évolutionniste), automatiques, universelles, jouant un rôle dans la survie reliées au système limbique: peur, colère, joie, tristesse, dégoût, surprise (expectative, acceptation). D'autres auteurs citent : « anger, despair, disgust, doubt, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise and worry » (Devillers *et al.*, 2003b). (Shafran Mohri, 2005) étudie la jovialité, la peur, la crainte et la satisfaction. Ce n'est pas une catégorisation courante car pour la plupart, le choix consiste en l'étude des émotions neutres, tristes et joyeuses.(Jiang *et al.*, 2000). A ce trio, certains rajoutent la peur et la colère.

Ainsi, (Bänziger, 2004) analyse la neutralité, la joie, l'ennui, la colère, la tristesse, la peur et l'indignation. (Devillers *et al.*, 2003), (Devillers *et al.*, 2003b), (Devillers *et al.*, 2003a), ont choisit les attitudes suivante: colère, peur, satisfaction, excuse et attitude neutre. La norme MPEG-4 possède aussi une nomenclature des affects (de Mareüil *et al.*, 2000); Elle se compose de : Colère, dégoût, peur, joie, surprise et tristesse. Cette normalisation vient du désir de synthétiser du texte par une voix émotive tout en synchronisant le son avec des mouvements faciaux d'un visage virtuel (PROJET INTERFACE). Chaque catégorie comporte des descripteurs continus sur deux ou trois dimensions (Schlossberg, 1954):

- positif/négatif, agréable/désagréable (évaluation)
- puissance/impuissance, tension/relaxation (puissance)
- activation/calme (activité)

D'après (Oudeyer, 2002), à l'opposé de la reconnaissance automatique des émotions par l'expression faciale (Samal Iyengar, 1992), les recherches sur la voix parlée sont encore très jeunes (Bosh, 2000). Pour exemple, le tableau suivant présente les caractéristiques des réponses faciales pour les affects primaires:

Affects primaires	Réponses faciales
Joie ou Jouissance	Sourire
Intérêt ou Excitation	Les sourcils sont baissés et le regard est fixé ou suit un objet.
Surprise ou Étonnement	Les sourcils sont relevés et les yeux clignotent.
Détresse ou Angoisse	Cri ou larmes
Colère ou Fureur	La mâchoire est serrée et le visage devient rouge.
Honte ou Humiliation	Les yeux et la tête sont baissés.
Mépris	La lèvre supérieure est relevée dans un sourire.
Dégoût	La lèvre inférieure est baissée et avancée.
Peur ou Terreur	Les yeux sont ronds et figés et le regard fixé ou s'éloignant de l'objet de la peur.

Neuf affects primaires et leurs réponses faciales, selon (Tomkins, 1984).

Les premières études ((Murray, 1993), (Williams Stevens, 1972)) n'avaient pas pour but de réaliser un système de reconnaissance efficace, mais plutôt de rechercher des corrélas qualitatifs généraux entre les paramètres acoustiques de la voix et les émotions qu'elle exprime (Lieberman Michaels, 1962). Par exemple, la joie tend à faire augmenter la moyenne de la fréquence fondamentale.

I.2.Applications:

Plus récemment, le besoin industriel d'un langage computationnel affectif (Picard, 1997) pousse la recherche à mettre en oeuvre des systèmes performants dans la reconnaissance des émotions (Bosh, 2000). On peut citer entre autres applications :

- Amélioration des systèmes de génération de parole à partir de texte,
- Agents assistants :
    - ° Adaptation au profil et état émotionnel d'utilisateur
    - ° Apprendre à ne pas gêner !
- E-mail expressif,
- Systèmes d'exploitation, interfaces ,
- Internet, bases d'images (mémorisation, stockage, récupération),
- Systèmes d'aide à l'enseignement, personnages animés, jeux,
- Compréhension et thérapie des troubles cognitifs (autisme),
- Robots (seuls, en groupe, ou en interaction avec humains),
- Applications artistiques et aux spectacles.
- norme MPEG-4 pour l'anglais, le français et l'espagnol (de Mareüil *et al.*, 2000): Projet INTERFACE avec des marqueurs du type <sadness>: Colère, Dégoût, Peur, Joie, Surprise, Tristesse. (Projet de langage textuel commun à la synthèse vocale et faciale)
- Une étude sur la détection des émotions dans des dialogues a été entreprise dans le cadre du projet européen AMITIÉS (call centers).
- Le brevet américain déposé par (Petrushin, 1999), vise à fournir à l'utilisateur d'un téléphone un retour sur son état émotionnel.

Conséquemment à ces besoins, un certain nombre de travaux sont consacrés à l'étude des émotions, de l'expressivité. Par exemple, le sujet de la thèse de C. Clavel est : « Analyse et détection des manifestations acoustiques d'états émotionnels liés à la peur » (Vasilescu *et al.*, 2004). On peut aussi citer des projets Européens, tels que EmoTV1, Amitiés Project, le réseau (Noe) HUMAINE (« Theories and Models of Emotion »), SpeechEmotion 2000, etc. Notons que le projet HUMAINE s'intéresse à l'expressivité dans le contenu audio et visuel, à l'analyse des gestes expressifs et à la performance artistique interactive, « Interactive Artistic Performance Testbed », en particulier dans la danse (« as artistic expression of human movement »).

I.3.Enjeu du stage :

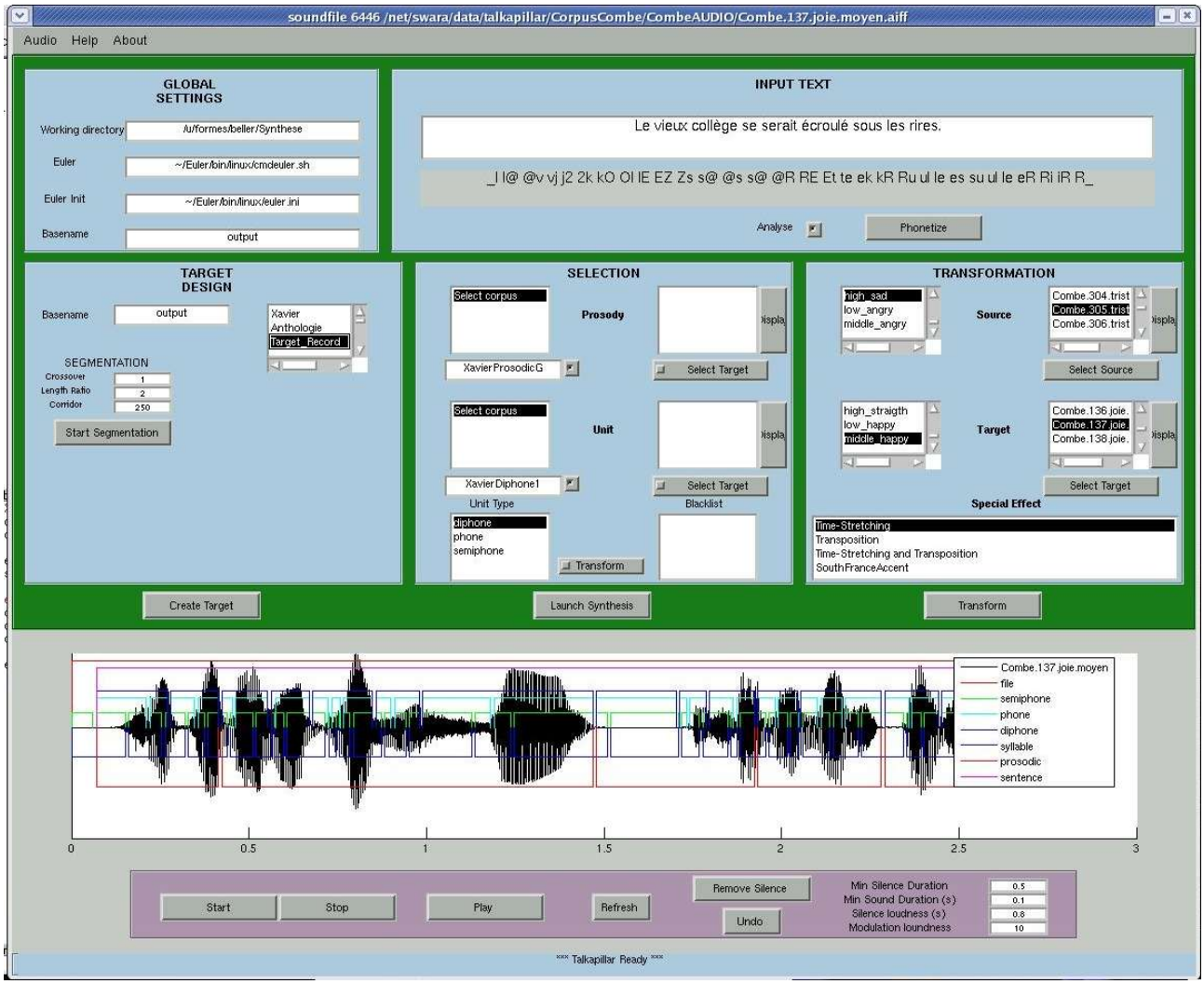
L'un des buts de ce stage est de comprendre la façon dont l'expressivité se traduit par la modulation de la prosodie et des caractéristiques acoustiques de la voix en général (qualité vocale), mais se distingue nettement des recherches citées ci-dessus. L'enjeu proposé sera, entre autres de répondre aux besoins des compositeurs et metteurs en scène pour le théâtre, le cinéma, la télévision et le multimédia en général. Il s'agit de pouvoir faire « prononcer » un texte par un générateur de parole avec la possibilité pour l'utilisateur d'indiquer et d'obtenir l'expressivité qu'il souhaite. Toutes proportions gardées, c'est ce qu'apporte l'interprétation d'un acteur à la prononciation d'un texte.

Le premier aboutissement de ce stage est donc une meilleure connaissance objective des corrélations entre l'expressivité dans la parole et ses descripteurs acoustiques. Le deuxième est la conception et la validation d'un modèle génératif de l'expressivité orienté vers le multimédia et la création artistique. Enfin une application permettra de synthétiser de la parole expressive à partir de descripteurs de haut niveau, relatifs aux catégories discrètes décrites précédemment. Une implémentation en temps réel peut-être envisagée.

I.4.Cadre de travail: Synthèse TTS concaténative

Un synthétiseur vocal par sélection d'unités (Beller, 2005) est en cours de développement à l'Ircam pour les applications artistiques mentionnée précédemment. Dans un stage de maîtrise, l'utilisation de patrons prosodiques réels a été étudiée et appliquée avec succès (Beller, 2004). La parole ainsi générée est intelligible et naturelle grâce à la concaténation de ces patrons prosodiques réels. Ce projet outil est continué durant ce stage de MASTER-2: nous souhaitons développer, plus avant, la part émotive et l'expressivité de l'intonation et des caractéristiques acoustiques dans notre modèle. Voici l'interface graphique (Juillet 2005) réalisée avec MATLAB du système offrant plusieurs modules:

- Global settings: Répertoire de travail, nom de la synthèse
- Input Text: Texte à synthétiser et visualisation de la phonétisation d'EULER
- Target Design: Pour entrer une phrase dans la base de donnée et s'en servir en tant que cible
- Selection: Sélection des corpus utilisé pour la synthèse
- Transformation: choix de la source, de la cible et de la transformation à utiliser.
- Visualisation de la forme d'onde et des unités. Réduction du temps de silence au début et à la fin.



Interface de TALKAPILLAR réalisée avec Matlab

L’intérêt de la méthode est la création, l’utilisation et l’adaptation d’outils dédiés à la parole et pouvant être exploités à des fins musicales et artistiques (Beller, 2005). Le cadre de travail de la synthèse par sélection d'unités est utilisé par de nombreuses équipes travaillant sur la synthèse de la parole émotive. Ainsi (Bulut *et al.*, 2002), (Lida *et al.*, 2003) concatènent des diphones extraits de bases de données émotives. (Black, 2003), l'un des pionniers de la synthèse concaténative fait de même en enregistrant plusieurs corpus émotionnels. (Tsuzuki *et al.*, 2004) limitent la taille de la base de données en entraînant une Chaîne de Markov Cachée (HMM) sur le corpus. Ainsi ils peuvent transformer des phrases issues de la sélection d'unité pour leur conférer des émotions modélisées par des processus HMM. Enfin, le brevet américain (Henton, 1997) décrit ci-dessous les commandes manuelles d'un synthétiseur TTS émotionnel:

Paramètres	**Commandes du Synthétiseur Vocal**
f0 moyen	BaselinePitch (pbas)
f0 variance	Pitch Modulation (pmod)
Débit	Speaking rate (rate)
Intensité	Volume (volm)
Pause	Silence (slnc)
Contour de f0	Pitch rise (/), pitch fall (\)
Durée	Lengthen (>), shorten (<)

Commandes d’un synthétiseur vocal expressif (Henton, 1997)

Il est important de souligner que nous recherchons avant tout à créer des voix émotives sans autres indices que l'émotion voulue: C’est à dire une synthèse TTS automatique.

Pour une alternative à la synthèse concaténative, voir (Murray *et al.*, 2000) qui trouve la synthèse par formants plus « maîtrisable » dans le sens ou l'on peut directement étudier l'influence d'un paramètre sur le résultat.

I.5.Plan de travail

La première partie consiste à construire une base de données en enregistrant un acteur récitant un texte avec des expressivités définies, dans une chambre anéchoïque. Puis à extraire les paramètres acoustiques de ces signaux grâce aux outils développés à l'IRCAM (segmentation temporelle, extraction de la fréquence fondamentale, etc...). On essayera d'analyser le substrat préverbal propre à l'expression acoustique des émotions (Auchlin Simon, 2004).

Une fois la base de données réalisée, nous allons tenter de mettre en place des descripteurs de haut niveau caractérisant les affects, de manière heuristique. Ainsi, ce rapport bibliographique s'est attaché à recenser à travers la littérature les différents descripteurs acoustiques utilisés dans la reconnaissance et la synthèse des émotions.

De manière à explorer la pertinence de ces descripteurs, nous essayerons de transformer le signal, soit par des techniques de morphing, soit par la recherche d'unités dans une base de donnée.

Trois méthodes peuvent alors être employées pour valider ces descripteurs:

- La première est la mise en place de tests perceptifs sur un nombre statistiquement raisonnable d'individus qui vont écouter des phrases et les classer par affects. Parmi ces phrases, on adjoint des phrases synthétisées dans les quelles on a fait varier des paramètres qui nous semblent valables. C'est le paradigme de l'analyse par la synthèse (Scherer, 2003).
- La seconde méthode est celle que nous utiliserons pour l'instant. C'est le tri sur la base du jugement personnel. Une démarche qui n'est pas tellement éloignée de nos applications artistiques.
- La troisième est la plus plébiscitée dans les articles car elle permet des comparaisons objectives. C'est l'analyse automatique d'un corpus grâce à des algorithmes de classification de données. Détecter des émotions dans la parole peut être vu comme une tache de classification qui consiste à assigner une catégorie émotionnelle à un signal de parole (Shafran Mohri, 2005). Cette classification peut s'effectuer par des algorithmes d'apprentissage exercés sur des bases de données. Ainsi, (Yacoub, 2003) utilise des réseaux de neurones, des machines à vecteurs de support, des K plus proches voisins et des arbres de décision. Pour un exemple plus détaillé on peut citer, (Oudeyer, 2003) qui utilise aussi des réseaux bayésiens naïfs, des mixtures de modèles gaussiens, la régression linéaire et d'autres classificateurs que l'on peut présenter de manière générale de la façon suivante:

- **Apprentissage supervisé**: Support Vector Machines, réseaux de neurones (Pereira, 2000), arbres de décision.
- **Apprentissage non supervisé**: Mixture gaussienne, Réseau Bayésien.

Ces différents algorithmes ne seront pas implémenter dans ce stage, faute de temps. De plus il est préférable de regarder ce qui se passe au niveaux des descripteurs « à la main » avant de les lancer dans une boite noire qui nous dira lesquels sont les meilleurs. Leurs implémentations suivront très probablement lors de la thèse. Ceci nous permettra d'extraire du corpus les meilleures corrélations de paramètres acoustiques relevant de telles ou telles émotions. Ce travail a déjà été esquissé (Oudeyer, 2002) mais pour l'anglais seulement. Rien de comparable ne semble à ce jour avoir été réalisé dans le cas du français parlé.

II.La prosodie:

II.1.Rappels:

« L'intonation joue des rôles multiples dans le langage de tous les jours. Elle reflète la structure hiérarchique de la phrase, et au-delà de la phrase, celle du discours. Elle distingue une question d'une réponse. Elle « désambiguïse » des séquences telle que « Je ne veux pas mourir idiot » (la prosodie doit préciser lequel des deux est l'idiot, le locuteur ou l'interlocuteur). Elle exprime des attitudes, des émotions. Elle n'arrive pas, cependant, à représenter des objets, des structures, des événements. Elle n'a pas de fonction représentative, comme les mots, et pas même une fonction figurative, non conceptuelle comme les gestes. » (Fónagy, 1983).

« Un accent expressif ou émotif renseigne sur l'état d'esprit du locuteur. Plusieurs nuances de sens peuvent être par là véhiculées. Les déplacements accentuels servent ainsi à transmettre des sentiments d'impatience, de colère, de doute, d'incertitude, d'amour ou de haine, etc. C'est le domaine de l'expression des émotions. » (Galarneau *et al.*)

La structure prosodique résulte d'interactions complexes entre différents niveaux d'organisation sémantico pragmatique, syntaxique et rythmique. Elle se manifeste par le jeu simultané de plusieurs paramètres acoustiques : la fréquence fondamentale Fo, le timbre, l'intensité, la durée des phonèmes. Perceptivement, la hauteur et son évolution, le rythme et le tempo (débit), le registre et le timbre mais aussi les pauses et les silences nous permettent la compréhension d'informations au-delà des mots prononcés. C'est cette deuxième partie du **double codage de la parole** (Fónagy, 1983)), qui lui confère un caractère "naturel" et évite la monotonie. Elle permet entre autre de véhiculer des informations ectolinguistiques ou phonostylistiques (expressivité, sentiments), de lever des ambiguïtés de sens entre deux phrases phonétiquement similaires et de structurer l'énoncé.

Dans le brevet américain déposé par (Petrushin, 1999), on peut lire que la hauteur est considérée comme l'indice le plus important pour la reconnaissance des émotions. Le registre couvert par la plupart des locuteurs est souvent divisible en 4 niveaux perceptivement distinguables : Nous les nommerons:

H+H+ : niveau le plus haut

HH

LL

L-L- : niveau le plus bas

La fréquence fondamentale f0 évolue dans ce registre. Son évolution au cours du temps décrit des contours. Une phrase est généralement composée d'une suite de contours qui ne suivent pas nécessairement la même orientation de pente. On observe cependant une déclinaison générale qui correspond à un abaissement de f0 du début à la fin de l'énoncé. La hauteur la plus basse correspond donc à la fin de cet énoncé et constitue ainsi un bon indice de segmentation. Ce phénomène à priori universel est de nature physiologique, mais il est géré par le locuteur à des fins linguistiques; il permet de délimiter la fin d'une phrase syntaxique. Il faut remarquer que l'on ne peut évaluer cette fréquence fondamentale que sur les segments voisés (voyelles et quelques consonnes...). Aussi, nous extrapolons celle-ci durant les segments non voisées afin d'avoir des contours continus. De plus, La hauteur de la voix étant fondamentalement différente selon le locuteur, on ne peut associer aux niveaux décrits précédemment des valeurs de fréquence fixes. De manière à comparer les groupes prosodiques entre eux, les contours de f0 sont normalisés en durée et en hauteur (Klabbers Santen, 2002).

II.2.Fréquence fondamentale f0:

La mesure de la fréquence fondamentale est donnée par l'algorithme YIN basé sur le principe de l'autocorrélation. (De Cheveigné et al., 2003). L'estimation est accompagnée d'une mesure de l'énergie et d'une mesure de l'apériodicité du signal (valeur comprise entre 0 et 1). Nous utilisons cette dernière pour définir un indice de confiance sur l'estimation de f0. Lorsque l'apériodicité est inférieur à un seuil (0,2), le signal est considéré comme apériodique et la mesure de f0 est alors évincée. Cela veut dire qu'en sortie, nous obtenons une courbe par morceaux de la fréquence fondamentale estimée que sur les segments voisés. Afin de débruiter la courbe des erreurs ponctuelles d'estimation:

- 2 ou 3 valeurs dans une région clairement non voisée
- Sauts d'octave

Nous filtrons la courbe de f0 par un filtre médian d'horizon comprenant 15 valeurs. Cela peut paraître un peu large comme fenêtrage temporel, mais nous souhaitons au final obtenir une courbe lisse appelée « contour de f0 » qui traduit plus une allure qu'une variation instantanée (de manière à ne pas prendre en compte le contexte phonétique). Dans le cadre du synthétiseur TTS, nous interpolons la courbe sur les segments non voisés afin d'avoir une valeur de f0 sur toute la phrase (voir (Beller, 2004) pour plus de détails).

II.3.Débit de parole:

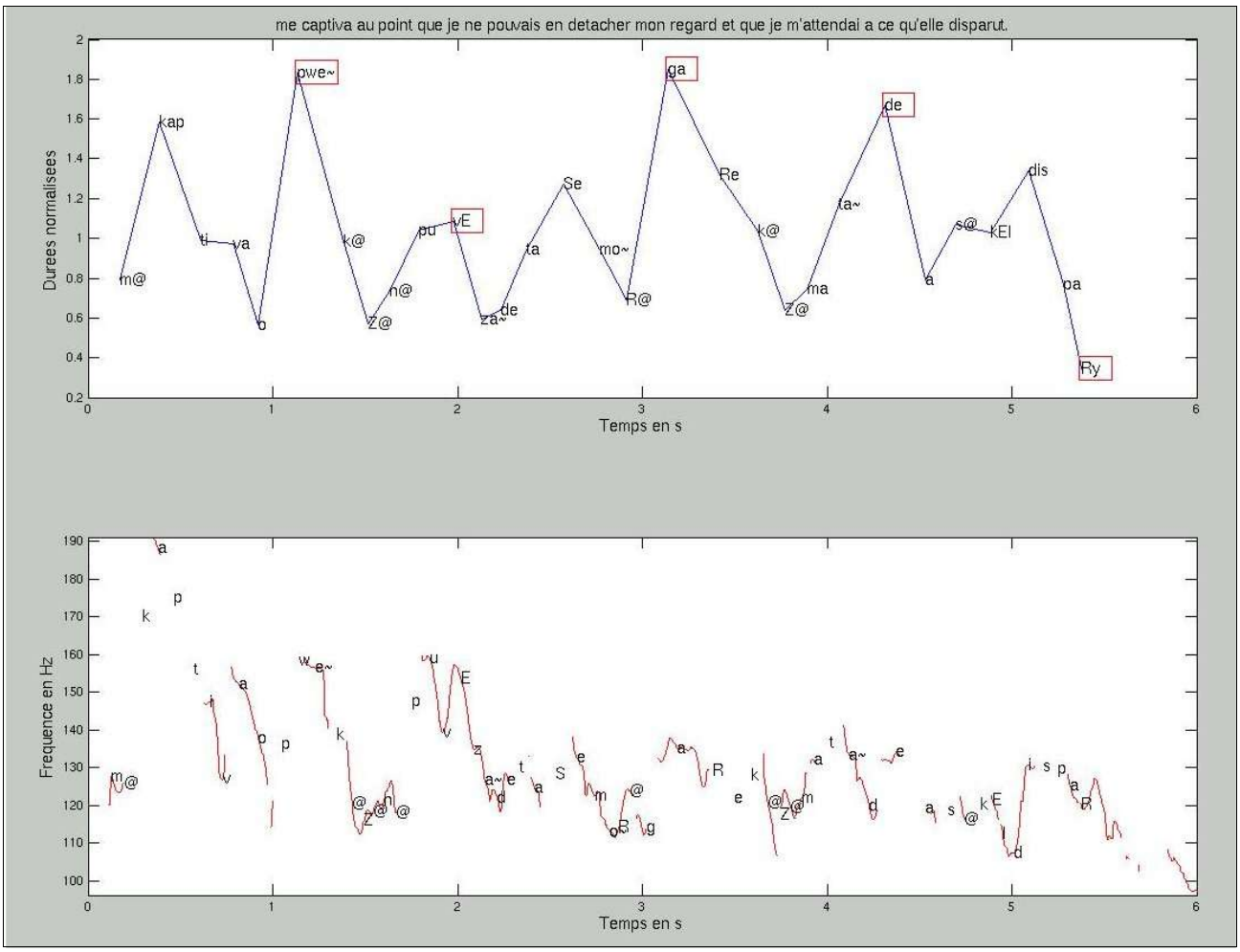
Contrairement à la fréquence fondamentale, le débit est une notion difficilement mesurable. Certainement car il n'y a pas de consensus au niveau de sa définition. P. Martin (Galarneau *et al.)* le définit comme un mouvement d'ensemble de l'énoncé. « Le débit porte sur tout ce qui est inclus dans un tour de parole, y compris les pauses silencieuses. Il se calcule en syllabes, en segments ou en mots. On distingue trois types de débits réguliers (lent, moyen, rapide) et deux types de changements de débit (accélération et ralentissement). Bien entendu, la durée des segments est affaire d'organisation temporelle. Mais la continuité des énoncés, en termes de fluidité ou d'hésitation, de même que le débit, en tant que tempo (cadence) des énoncés, le sont aussi. Les pauses peuvent être remplies (par des phatèmes, par exemple), ou silencieuses. La parole peut donc être continue ou interrompue. En matière de débit, on peut distinguer entre la vitesse d'articulation d'unités comme la syllabe, à savoir le débit articulatoire, puis le débit de la parole, qui comprend les hésitations, les interruptions et les pauses. En anglais, le débit articulatoire moyen est d'environ 5,3 syllabes par seconde, alors que le débit moyen de parole est d'un peu plus de 200 mots à la minute. »

On parle généralement d'une langue à cadence syllabique quand celle-ci ne génère pas son rythme sur la base d'un véritable accent de mot, l'emphase et l'expressivité étant mises à part. Le rythme de la langue s'appuie plutôt sur la simple récurrence des syllabes, qui tendent alors à avoir une durée à peu près égale. Ceci suppose naturellement que le timbre vocalique se maintienne hors accent. **Le français est un bon exemple de langue à cadence syllabique**. « La normalisation au débit semble s'effectuer en grande partie intra syllabiquement » (Miller, 1987). Une démonstration assez convaincante de l'existence d'un syllabaire pour la perception serait l'existence d'un effet de fréquence syllabique, comme celui que (Wheeldon, 1994) ont trouvé en production. Toutefois, il faudrait décorréler la fréquence des syllabes de celles des diphones et des demi syllabes qu'elles contiennent, tâche pratiquement impossible à réaliser. Car il existe des syllabes contenant 5 phonèmes et d'autres n'en contenant que 2. Ainsi, on peut présager que ces dernières dureront moins longtemps. Toutefois, nous faisons l'hypothèse que si une syllabe est allongée par rapport à une autre, c'est pour une raison prosodique et non phonétique. « Syllabes et phonèmes constituent les meilleurs candidats pour évaluer le débit de parole. Celui-ci est d'ailleurs plus corrélé au débit syllabique qu'au débit phonétique » (Rouas *et al.*, 1998). « Dans la mesure ou s'est établi un consensus sur la réalité psycho rythmique de la syllabe en Français » (Zellner, 1998), nous basons notre modèle de débit sur la variation des durées des syllabes.

De manière à segmenter l'énoncé en syllabes, phones, diphones et semiphones, nous utilisons l'outil d'alignement avec le texte développé à l'IRCAM (Beller, 2004). En synchronisant la

segmentation temporelle à la description phonétique, grammaticale et lexicale d'EULER (Dutoit et al., EULER), on parvient non seulement à une description symbolique de ces segments, mais à un étiquetage phonétique et syllabique qui nous permet de construire une courbe de débit. Cette courbe est décrite par la durée des syllabes et croisse donc pour un ralentissement. Enfin on la normalise par la moyenne des durées des syllabes de la phrase pour que l'accélération ou le ralentissement soit relatif au débit moyen de la phrase. Cette moyenne nous renseigne sur le débit global de la phrase.

Ces analyses sont illustrées dans la figure suivante:



Débit normalisé et évolution de la fréquence fondamentale d'un signal de parole

On y distingue plusieurs informations:

- La courbe de débit (en haut) présente des maxima locaux relatifs à des ralentissements.
- Les syllabes encadrées en rouge sont des accents finaux de groupes prosodiques générés par le modèle prosodique d'EULER. On voit que pour cette phrase, le modèle concorde à la prononciation du locuteur.
- La courbe de la fréquence fondamentale (en bas) n'est pas celle que nous utilisons puisqu'on l'interpole entre les parties voisées. Ici ne figurent que les régions décrites comme voisées par le seuillage sur l'apériodicité. On y observe la déclinaison générale due à l'expiration (moins d'air dans les poumons en fin de phrase) et quelques variations intrinsèques des phones dues à la prononciation du locuteur. Les parties non voisées correspondent bien au consonnes (plosives, sifflantes, fricatives...)

Le principal avantage de cette méthode d'estimation du débit, est qu'elle fournit une courbe relativement instantanée. En effet, de nombreuses études sur la parole émotionnelle utilise un indice de débit global sur toute la phrase voire même sur tout un énoncé (Chung, 2000). Celui-ci est souvent mesuré en syllabes par seconde.

II.4.Énergie:

L'estimation de l'énergie nous est donnée par l'analyse YIN. Cependant, celle-ci ne se révèle pas d'elle-même, être un bon indice de l'intensité prosodique car elle est très dépendante du contexte phonétique. En effet, une voyelle peut s'avérer être beaucoup plus énergétique qu'une plosive (présence d'un silence avant l'explosion), alors qu'aucun changement d'intensité ne sera perçu par l'auditeur. Pour obtenir un indice de la variation d'intensité perceptive au cours de la phrase, nous faisons une moyenne statistique des moyennes des énergies par classes de segments. Par exemple, on prend tous les phones « E » de notre base de données, et on fait la moyenne. Si la moyenne de l'énergie d'un « E » et au-dessus de la moyenne des énergies de sa classe, cela veut dire que ce « E » aura été prononcé avec une emphase d'intensité. Cette variation correspond alors à une stratégie prosodique indépendante du contexte phonétique.

Nous calculons d'autres caractéristiques statistiques pour chaque segment:

- La moyenne des f0 pour chaque classe
- l'écart en fréquence fondamentale de chaque segment par rapport à la moyenne des f0 de sa classe
- La moyenne des durées pour chaque classe
- l'écart en durée de chaque segment par rapport à la moyenne des durées de sa classe

Ces descripteurs reposent sur l'hypothèse que l'individu ne s'exprime pas par des valeurs absolues mais plutôt par des distorsions ou variations de valeurs habituelles.

II.5.Qualité Vocale:

(Bulut *et al.*, 2002) indique très clairement que la prosodie ne suffit pas à décrire les émotions, il manque : la qualité vocale. Ce terme désigne toute l'identité de la voix à travers son timbre. Comme pour la musique, ce champ d'investigation est relativement récent et se révèle de plus en plus important pour la caractérisation d'un événement sonore. En ce qui concerne les émotions, la qualité vocale est souvent traduite par des mots métaphoriques: « soufflé », « crispé », « détendu »... Qui découlent directement de l'état physiologique de la personne. Là aussi comme en musique, il n'existe pas encore de consensus pour la dénomination précise du timbre car celui-ci possède un domaine de variabilité quasiment infini. Nous allons voir dans la partie suivante ou l'on recense les différents descripteurs acoustiques présentés dans la littérature, au combien les descripteurs de la qualité vocale sont peu nombreux face aux descripteurs prosodiques.

III.Base de données:

III.1.Émotions vécues ou simulées: (Chung, 2000)

III.1.a.Émotions vécues:

Lorsqu'il s'agit d'étudier des émotions, il faut faire une distinction de part la nature du contexte dans lesquelles elles ont été exprimées. En effet, dans l'étude de l'émotion vocale, le corpus d'émotions vécues se constitue des émotions spontanées, produites par la transformation des états d'âme du sujet et enregistrées avec le moins d'intervention possible de l'expérimentateur. Ce genre de corpus reflète les meilleures caractéristiques de l'émotion naturelle mais, dans des expériences phonétiques et psychologiques, son utilisation est assez limitée par des problèmes méthodologiques. L'étude de l'émotion se heurte à un dilemme similaire à celui de l'étude de la parole spontanée: le but de l'étude est d'analyser les émotions exprimées dans un milieu naturel mais l'acquisition et l'expérimentation de ce genre de données sont extrêmement difficiles. C'est ce que Labov (Labov, 1972) appelle le paradoxe de l'observateur. Selon l'expression de Fagyal (Fagyal, 1995), l'authenticité des données et le contrôle expérimental sont liés par une fonction inverse. Plus les données de l'émotion naturelle sont authentiques, moins les contrôles sont disponibles à l'expérimentateur sur l'organisation et la manipulation des données.

Diverses méthodes sont inventées pour faire un compromis entre l'authenticité des données émotionnelles et le contrôle expérimental. La provocation attendue, souvent utilisée par les psychologues, consiste en la provocation des émotions bien définies dans un cadre de laboratoire. L'expérimentateur présente aux sujets des stimuli qui sont censés provoquer certaines réactions émotionnelles (des images, des morceaux de musique ou des textes littéraires) et il enregistre leurs réactions exprimées dans la voix pour comparer la voix émotionnelle avec la voix non émotionnelle (Skinner, 1935). Cette méthode permet un grand contrôle sur la production des émotions et une bonne acoustique d'enregistrement mais la provocation des émotions par des moyens artificiels peut être problématique du point de vue psychologique. La provocation inattendue de l'émotion se trouve dans un milieu naturel, hors du laboratoire, et dont la procédure est la même que la provocation attendue, sauf que les sujets ne savent pas qu'ils sont enregistrés. Dans l'expérience de (Bonner, 1943), un signal d'alarme est présenté en plein cours, ce qui rend les étudiants anxieux et stressés. Leurs voix émotionnelles sont enregistrées discrètement et ces expressions naturelles sont comparées avec les expressions des émotions simulées par les mêmes étudiants une semaine après. Le résultat montre qu'il y a des changements prosodiques sous la tension émotionnelle mais la nature des changements varie selon les individus. Par exemple, certaines voix montre l'augmentation du f0 moyen dans l'état anxieux mais d'autres voix montrent la diminution du f0 moyen dans la même situation.

Les émotions exprimées dans des interactions sociales sont plus naturelles que les émotions provoquées par des stimuli, reflétant mieux la communication des émotions naturelles. Elles sont produites dans des milieux quotidiens comme la maison, l'école et le lieu de travail, avec une intensité modérée (voir (Hutter, 1968), (Léon, 1970)). Les expressions émotionnelles dans une série de conversations entre un employé et son supérieur d'une compagnie d'électricité sont étudiées par (Streeter *et al.*, 1983). Les auteurs constatent que le f0 moyen et le niveau d'amplitude augmentent en fonction de la difficulté des tâches.

III.1.b.Émotions simulées:

Les émotions simulées sont souvent utilisées dans la recherche de l'expression émotionnelle, en raison de la facilité de la manipulation systématique des données. Un des sujets majeurs dans l'étude de l'émotion vocale est de savoir comment les différentes émotions sont identifiées à travers les indices prosodiques, non verbaux. Pour expérimenter ce genre de question, les chercheurs construisent les données par la simulation des expressions émotionnelles avec l'acteur ou la machine

de synthèse vocale. ((Fairbanks Pronovost, 1938), (Fónagy Bérard, 1972), (Scherer Oshinsky, 1977), (Laukkanen *et al.*, 1996).

En ce qui concerne la simulation par l'acteur, professionnel ou non, le chercheur détermine d'abord les émotions à étudier, et puis décrit au locuteur le contexte de chaque émotion, qui est censé provoquer une telle émotion. Le locuteur imite les différentes émotions, en modifiant des traits prosodiques (intonatifs) de la voix. Le contenu verbal est contrôlé dans le sens qu'il est soit constant à travers les différentes émotions soit éliminé au moyen de filtrage pour qu'il ne joue pas de rôle dans l'expression ou dans la perception de l'émotion. Les données ainsi acquises facilitent l'analyse comparative, grâce à son organisation systématique. La comparaison porte sur la variation des paramètres prosodiques en fonction des émotions. Ce genre de données est d'un haut degré de contrôle mais peu spontané. Vu que l'enregistrement a lieu dans un studio insonorisé, la qualité acoustique est généralement bonne, ce qui est essentiel pour les mesures exactes des traits prosodiques.

Certains chercheurs de l'expression émotionnelle préfèrent les données d'émotions simulées par des locuteurs non professionnels à celles simulées par des acteurs professionnels ((Kaiser, 1962), (Bezooijen, 1984), (Leinonen *et al.*, 1997)). Étant conscients des effets spéciaux dus au jeu théâtral dans la voix professionnelle, ils considèrent l'expression émotionnelle non professionnelle comme plus naturelle que l'expression professionnelle. Quand on demande au locuteur non professionnel de simuler des émotions sans instruction spécifique sur la façon de s'exprimer (voir la Figure 13), il se réfère à ses propres réactions émotionnelles dans le passé, d'où l'authenticité de son expression émotionnelle.

Dans le cadre de notre travail, la spontanéité des émotions n'est pas le facteur d'importance puisque nous cherchons plutôt à synthétiser des affects de la même manière que le ferait un acteur. De plus, nous nécessitons un protocole d'observation assez contraignant de manière à exploiter proprement les données. De ce fait nous avons opté pour la constitution d'une base de données à partir de phrases énoncées par un acteur professionnel. Nous recherchons les traces acoustiques issues de processus aussi bien cognitifs que physiologiques. Au fond, nous cherchons à reproduire les indices acoustiques produits par un acteur professionnel lors d'une performance et qui semblent pertinents du point de vue de notre perception.

III.2.Bases de Données existantes:

On recense aujourd'hui de nombreuses bases de données de paroles émotionnelles dans de nombreuses langues et toujours orientées vers une application précise. Celles-ci se distinguent par les catégories d'émotions qu'elles regroupent, par leurs tailles, par la nature du matériau (discours, phrases, conversations; émotions vécues ou simulées)... La plupart sont en anglais ou chinois. Aussi, peu de bases de données en Français existent. C'est pour cela que nous avons émis le souhait d'en constituer une. Les méthodologies existantes ont permis l'établissement d'un protocole d'enregistrement. Nous y reviendrons dans le prochain chapitre concernant l'établissement de ladite base de donnée.

La distinction la plus forte entre les bases de données, après la langue, provient de la manière dont sont générés les extraits sonores de parole émotionnelle. Les différentes méthodologies utilisées sont toujours issues d'un choix très clair au départ: Vouloir réunir des émotions vécues ou bien simulées? Ainsi de nombreuses bases de données regroupent des extraits sonores tirées de situations réelles dans lesquelles les émotions se sont manifestées spontanément. Par exemple, (Arnfield and al, 1995) extraient différentes émotions tirées d'interviews radio/télévision dans lesquelles les participants sont conviés à revivre des expériences émotives intenses (corpus en anglais). D'autres sont construites à partir de mises en situation et donc d'émotions induites. Ainsi, (Scherer, 1989) ont bâtit une base de données contenant deux types de parole stressée, en induisant ces deux états par la présentation de photos ('slides'): Un stress cognitif induit par des slides contenant des problèmes logiques; Un stress émotionnel induit par des images de corps humains blessés ou malades. D'autres étudient la dépression et l'état suicidaire grâce à des enregistrements de sessions thérapeutiques et de conversations téléphoniques.

Toutes ces bases de données ont l'avantage de regrouper des enregistrements d'émotions vécues, mais ont souvent l'inconvénient de présenter des données inhomogènes (différents locuteurs, différentes conditions d'enregistrement...).

Vis à vis de nos besoins, nous nous sommes donc tournés vers les bases de données d'émotions simulées ou actées. Nous détaillons dans le tableau suivant un recensement de diverses bases de données de parole émotionnelle selon la nomenclature proposée par le site HUMAINE (Schröder, 2001):

Identifiant	émotions	Taille	matériau	Langue
Danish Emotional Speech Database (Engberg et al., 1997)	Colère, joie, tristesse, surprise, neutre	4 sujets lisent 2 mots, 9 phrases, et 2 passages avec différentes émotions	Phrases non colorées émotivement	Danois
Groningen ELRA corpus number S0020 (www.icp.inpg.fr/ELRA)	Base orientée seulement partiellement vers les émotions	238 sujets lisent 2 textes courts	Phrases	Hollandais
Berlin database (Kienast & Sendlmeier 2000; Paeschke & Sendlmeier 2000) http://www.expressive-speech.net/	Colère forte, ennui, dégoût, peur panic, joie, tristesse douloureuse, neutre	10 sujets (5 hommes, 5 femmes) lisent 10 phrases chacun	Phrases sémantiquement neutres	Allemand
Pereira (Pereira, 2000)	Colère forte, colère froide, joie, tristesse, neutre	2 sujets lisent 2 expressions chacun	1 Phrase sémantiquement neutre, et 4 chiffres répétés	Anglais
van Bezooijen (van Bezooijen, 1984)	Colère, dégoût, contenu, peur, joie intéressée, tristesse, honte, surprise, neutre	8 (4 hommes, 4 femmes) lisent 4 phrases	Phrases sémantiquement neutres	Hollandais
Abelin (Abelin 2000)	Colère, dégoût, dominance, peur, joie, tristesse, surprise, timidité	1 sujet	Phrases sémantiquement neutres	Suédois
Yacoub (Yacoub 2003) (data from LDC, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28	Neutre, colère forte, colère froide, joie, tristesse, dégoût, panique, anxiété, désespoir, fierté, exaltation, intérêt, honte, ennui, mépris	2433 expressions de 8 acteurs	Phrases	Anglais

Comme on peut le voir dans la colonne langue du tableau ci-dessus, il n'a pas été répertorié de base en Français. C'est pour cela que nous avons choisi d'en constituer une.

III.3.Base de données créée:

La base données que nous avons constitué ressemble un peu à celle de Abelin (dans le tableau précédent), mais en français. Un seul acteur prononce un texte de 26 phrases sémantiquement neutres ou plutôt émotionnellement neutres de part leur signification:

1. bonjour.
2. comment.
3. au revoir.
4. assieds toi ici.
5. a toute a l'heure.
6. quelle heure est-il?
7. je ne reconnais pas cet endroit.
8. on t'a dit ça\, la veille au soir\, par téléphone.
9. tu m'obliges a parler\, devant tout ce monde.
10. l'idéal est de trouver\, un endroit assez élevé.
11. c'était bien avant\, que je ne trouve du travail.
12. la veille du départ\, ils ont dormi dans un hôtel.
13. quelque chose\, que tout le monde pourrait porter.
14. ce ne sont que des lumières\, et pas autre chose.
15. tu veux me parler du travail\, que j'avais avant d'être ici.
16. c'était dans l'enclos du fond\, que ça devait se passer.
17. une fois le rendez-vous fixe\, il est impératif de s'y tenir.
18. le mieux\, pour regarder ce spectacle\, c'est de s'asseoir.
19. un costume qui ne soit\, ni trop excentrique\, ni trop neutre.
20. Ca fait un quart de page\, avec une photo en noir et blanc.
21. tout est déjà organise\, même la presse a été convoquée.
22. dessous la photo\, il y avait une légende\, qui expliquait tout.
23. il faut préserver une certaine dignité\, autour de l'évènement.
24. c'était le même nom\, que la personne qui avait signe\, cet article.
25. j'ai lu dans le journal\, qu'on avait inaugure la statue\, dans le village.
26. il fallait prendre a gauche\, au rond point principal\, et ensuite\, filer tout droit.

Texte utilisé pour l'établissement de la base de donnée de parole émotionnelle

Ce texte a pour espoir d'être sémantiquement neutre: Nous espérons que ces phrases peuvent avoir un sens dans tous les contextes émotionnels utilisés afin de ne pas créer de décalages dus à leurs sens. Les mots soulignés sont les mots que l'acteur doit mettre en emphase quelque soit l'affect exprimé. Ceci afin d'éviter le plus possible des variations de sens et donc de prosodie qui ne relèverait pas directement de l'expression émotionnelle. Dans le même but, les groupes de sens ou groupes intonatifs ont été marqués à l'avance par des « \, ». L'acteur y fera des pauses silencieuses ou sonores mais devra respecter ces frontières pour chaque affect. Ce texte est répété plusieurs fois avec différentes émotions et plusieurs niveaux de valence (faible, moyen, fort) pour les émotions les plus importantes (ci-dessous en gras):

- **Droit (variation de f0 minimale demandée)**
- **Neutre**
- *Neutre interrogatif*
- **Joie**
- **Ennui**
- **Tristesse**
- **Colère**
- **Peur**
- *Indignation*
- *dégoût*
- *Surprise positive*
- *Surprise négative*

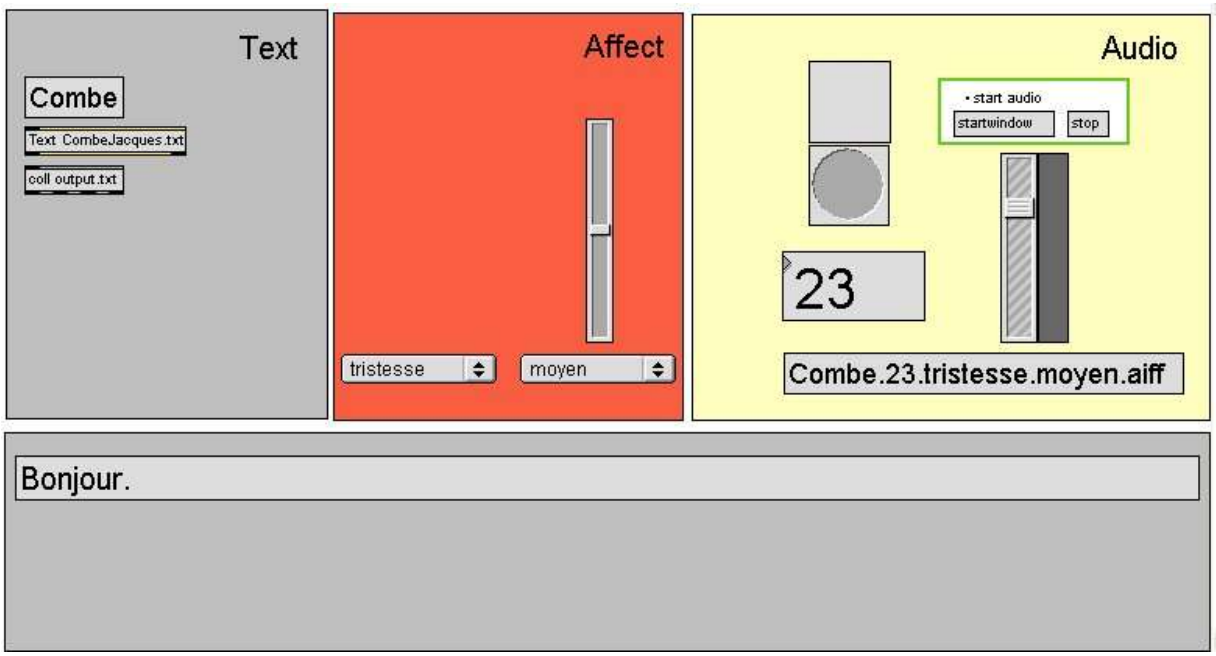
Émotions retenues pour l'expérience

L'IRCAM jouit d'une salle anéchoique qui permet des enregistrements de hautes qualités sans réverbération. Les signaux enregistrés dans ce type de salle se prêtent donc particulièrement bien à des opérations de segmentation. Ce point est primordial dans l'étude de la parole et a permis, notamment, de constituer des bases de données de parole pour le synthétiseur text-to-speech TALKAPILLAR développé ici. Nous avons donc procéder de la même manière et mis à profit la chambre sourde durant les enregistrements. La parole de l'acteur a été captée par un microphone statique de qualité, puis directement aquisitionné par une carte « EDIROL FA-101 » selon les caractéristiques suivantes:

- Fréquence d'échantillonnage : 44100 Hz.
- Numérisation: 16 Bits

Afin de faciliter la segmentation en phrases et la conduite de l'enregistrement, nous avons créer un programme appelé « recorder », grâce au langage graphique MAX/MSP. Ce programme permet à l'utilisateur de spécifier l'affect qu'il désire manifester, la phrase qu'il va dire et de déclencher/arrêter l'enregistrement en appuyant avec le pied sur un pédale MIDI reliée. Lorsque l'acteur appuie sur cette pédale, le programme écrit sur le disque dur de l'ordinateur portable relié à la carte d'acquisition les fichiers suivants:

- L'enregistrement du signal sous le format « .aiff ». Le nom donné au fichier est de la forme: « corpus.numéro.affect.valence.aiff » dans lequel le corpus correspond au nom de l'acteur (ou autre), le numéro est un index incrémenté à chaque enregistrement, l'affect fait parti d'une liste déterminée préalablement, la valence représente le degré d'intensité (faible, moyen, fort).
- Le fichier texte qui regroupe toutes les phrases prononcées auxquelles on adjoint le même index « numéro » de manière à établir une correspondance directe entre le fichier audio et le texte. Ceci facilite grandement les étapes ultérieures de segmentation par alignement et de description linguistique.



Interface du patch Recorder réalisé avec MAX/MSP pour faciliter l'enregistrement

L'enregistrement a eu lieu le mercredi 13 Avril 2005 et a duré environ quatre heures. A peu près 700 phrases ont été recueillies. Elles sont recensées dans le tableau suivant :

Affect	**Valence**	**Nombre de phrases**	**Nombre de phrases retenues**
Droit (variation de f0 minimale demandée)	Faible	26	22
	Moyen	26	26
	Fort	26	26
Neutre	moyen	52	48
Neutre interrogatif	moyen	26	18
Joie	Faible	26	25
	Moyen	26	26
	Fort	52	30

Affect	Valence	Nombre de phrases	Nombre de phrases retenues
Ennui	Faible	26	26
	Moyen	26	26
	Fort	26	26
Tristesse	Faible	26	26
	Moyen	26	22
	Fort	26	18
Colère	Faible	26	21
	Moyen	26	5
	Fort	26	4
Peur	Faible	26	22
	Moyen	26	16
	Fort	26	12
Indignation	Fort	26	18
Dégoût	Fort	26	12
Surprise positive	Moyen	26	18
	Fort	26	11
Surprise négative	Moyen	26	26
	Fort	26	9
TOTAL		702	539

La dernière colonne du tableau indique le nombre de phrases retenues par catégories pour les analyses ultérieures. En effet, des phrases ont été parfois écartées pour des raisons de mauvaise interprétation (chat dans la gorge, erreur de prononciation, affect jugé mal interprété par l'acteur lui-même) ou pour des raisons de mauvaise acquisition (mauvais RSB, bruits...). Au final, 539 phrases constituent une base de données propre et homogène.

Ces phrases dont nous possédons la représentation symbolique, C'est à dire le texte, sont ensuite alignées et segmentées grâce un outil de segmentation développé à l'IRCAM (Beller, 2004): De manière à segmenter l'énoncé en syllabes, phones, diphones et semiphones, nous utilisons l'outil d'alignement avec le texte développé à l'IRCAM (Beller, 2004). En synchronisant la segmentation temporelle à la description phonétique, grammaticale et lexicale d'EULER (Dutoit et al., EULER), on parvient à décrire symboliquement et donc à un étiqueter phonétiquement et syllabiquement les segments. Cela nous est très utile pour établir des classes de phonèmes en vue d'études statistiques par exemple.

L'acteur, Jacques Combe, a été retenu pour sa motivation et son expérience dans le domaine. Comédien et conteur de profession, il a déjà travaillé en studio pour créer des disques d'histoires et de contes et il est donc habitué à placer des affects et à les exprimer seulement par sa voix.

IV.Descripteurs acoustiques des émotions:

IV.1.État de l'art:

Dans cette partie, nous dressons un panorama de descripteurs acoustiques des émotions rencontrés dans la littérature. Diverses études ont mis en lumière la pertinence de l'extraction de données acoustiques du signal afin de le catégoriser par rapport à l'état émotif de son locuteur. Ce qui nous intéresse ici, est de répertorier ces différents descripteurs.

Cette partie nous permet aussi de voir quelles émotions sont jusqu'ici analysées et par quels indices, elles se sont fait remarquer par les algorithmes de classification. Ainsi, (Jiang *et al.*, 2000) écrit que « la tristesse peut-être caractérisée par une hauteur de moyenne basse, de faible variance et aux contours plats. Ces variations de hauteur limitées réalisent le ton de la syllabe et il y a une légère déclinaison dans l'intonation de la phrase. La joie possède un débit plus rapide et une déclinaison moins forte. La variance est quand à elle encore plus étroite ».

(Yacoub, 2003) nous permet d'introduire deux descripteurs assez répandus que sont le jitter et le shimmer. En effet il utilise les descripteurs acoustiques suivants:

- Fréquence Fondamentale:
  - minimum, maximum, moyenne et variance de f0
  - minimum, maximum, moyenne et variance de la dérivée de f0
  - minimum, maximum, moyenne et variance du jitter

Définition du Jitter: (Chung, 2000)

La perturbation de f0 (« jitter » en anglais) réfère à la micro variation du f0, c'est-à-dire la variation du f0, cycle par cycle, dans une unité acoustique. La mesure de jitter est proposée par (Lieberman, 1961) en tant que description quantitative de la qualité de la voix, surtout en ce qui concerne la voix émotionnelle. Il s'agit de la variation des périodes adjacentes dans un signal acoustique, dont l'estimation peut être accomplie à travers diverses méthodes d'analyse ((Horri, 1982), (Klingholz Martin, 1985), (Schoentgen de Guchtenneere, 1995)).

Ici, le jitter est mesuré par filtrage linéaire de la fréquence fondamentale:

$$H = 0.25*[-1, 3, -3, 1]$$

- Intensité:
  - minimum, maximum, moyenne et variance de l'enveloppe de l'énergie
  - minimum, maximum, moyenne et variance de la dérivée de l'enveloppe de l'énergie
  - minimum, maximum, moyenne et variance du shimmer (calculé comme le Jitter)

Définition du Shimmer: (Chung, 2000)

La perturbation d'intensité (« shimmer » en anglais) montre la variation de l'intensité, cycle par cycle, dans une unité acoustique. Le shimmer, comme le jitter, varie selon les différentes phonations et influence la perception de la qualité de la voix (comme le degré de rudesse vocale), mais la variation du shimmer est moins pertinente dans l'expression et la perception de l'émotion que la variation du jitter ((Horri, 1982), (Klingholz Martin, 1985)). Les valeurs du shimmer et du jitter sont trouvées corrélées dans l'expérience de (Takahashi Koike, 1975).

Ici, le shimmer est calculé de la même façon que le jitter, c'est à dire par filtrage linéaire de l'enveloppe de l'énergie.

- Débit:

Le débit est défini par (Yacoub, 2003) par la durée de segments audibles: Si une portion du signal est d'énergie plus faible que 1% de son maximum, alors cette portion de signal est considérée comme inaudible. Les segments inaudibles sont ainsi reliés aux pauses silencieuses. On calcule sur les segments audibles:

- minimum, maximum, moyenne et variance de la durée des segments audibles
- minimum, maximum, moyenne et variance de la durée des segments inaudibles
- Rapport entre les durées totales des segments audibles et des segments inaudibles
- Rapport entre la durée totale des segments audibles et la durée de la phrase
- Rapport entre la durée totale des segments inaudibles et la durée de la phrase
- Rapport entre la moyenne des durées des segments audibles et celle des segments inaudibles
- Le nombre de trames audibles divisées par le nombre de trames inaudibles

Il semble que le débit soit assez difficile à quantifier. Pour l'anglais, il est très souvent associer aux durées des segments voisés. Cette définition est préconisée par (Bänziger, 2004) qui a du passer par une segmentation manuelle au préalable (étape assez longue et fastidieuse qui ne permet pas l'étude de vastes corpus):

Code	Description
int.min	intensité minimale, mesurée en dB
int.max	intensité maximale, mesurée en dB
int.étdu	(Int_max-Int_min), étendue de l'intensité (dB)
int.moy	intensité moyenne, mesurée en dB
int.sd	écart-type de l'intensité, mesuré en dB
f0.min	minimum de la fréquence fondamentale (Hz)
f0.max	maximum de la fréquence fondamentale (Hz)
f0.étdu	(f0_max-f0_min), étendue de la f0 (Hz)
f0.moy	moyenne de la fréquence fondamentale (Hz)
f0.sd	écart-type de la fréquence fondamentale (Hz)
f0.05c	5ème centile de la fréquence fondamentale (Hz)
f0.25c	25ème centile de la fréquence fondamentale (Hz)
f0.50c	médiane de la fréquence fondamentale (Hz)
f0.75c	75ème centile de la fréquence fondamentale (Hz)
f0.95c	95ème centile de la fréquence fondamentale (Hz)
f0.05-95c	(f0_95c-f0_05c) étendue de la f0 (Hz)
dur.s/tot	durée des segments silencieux sur la durée totale
dur.n/tot	durée des segments non voisés sur la durée totale
dur.v/tot	durée des segments voisés sur la durée totale
dur.v/art	durée des segments voisés sur la durée articulée
dur.tot	durée totale des expressions (en secondes)
Pour les segments voisés	
v.125-200	% d'énergie entre 125 et 200 Hz
v.200-300	% d'énergie entre 200 et 300 Hz
v.300-500	% d'énergie entre 300 et 500 Hz

Code	Description
v.500-600	% d'énergie entre 500 et 600 Hz
v.600-800	% d'énergie entre 600 et 800 Hz
v.800-1k	% d'énergie entre 800 et 1000 Hz
v.1k-1.6k	% d'énergie entre 1000 et 1600 Hz
v.1.6k-5k	% d'énergie entre 1600 et 5000 Hz
v.5k-8k	% d'énergie entre 5000 et 8000 Hz
v.0-500	% d'énergie en dessous de 500 Hz
v.0-1k	% d'énergie en dessous de 1000 Hz
Hamm	Indice de Hammarberg: intensité max. entre 0 et 2 kHz moins intensité max. entre 2 et 5 kHz
Pour les segments non voisés	
n.125-250	% d'énergie entre 125 et 250 Hz
n.250-400	% d'énergie entre 250 et 400 Hz
n.400-500	% d'énergie entre 400 et 500 Hz
n.500-1k	% d'énergie entre 500 et 1000 Hz
n.1k-1.6k	% d'énergie entre 1000 et 1600 Hz
n.1.6k-2.5k	% d'énergie entre 1600 et 2500 Hz
n.2.5k-4k	% d'énergie entre 2500 et 4000 Hz
n.4k-5k	% d'énergie entre 4000 et 5000 Hz
n.5k-8k	% d'énergie entre 5000 et 8000 Hz
n.0-500	% d'énergie en dessous de 500 Hz
n.0-1k	% d'énergie en dessous de 1000 Hz

Codes et descriptifs des paramètres acoustiques utilisés par (Bänziger, 2004)

Dans la même idée, on peut citer le brevet américain (Petrushin, 1999) qui de manière à fournir à l'utilisateur d'un téléphone, un retour sur son état émotionnel, utilise les descripteurs suivants, dont la moyenne, la variance, le maximum, le minimum et l'écart entre ces deux dernières valeurs sont calculés:

- f0: On ajoute une interpolation linéaire sur les segments voisés afin d'obtenir un contour de hauteur.
- Énergie: On calcule un ratio d'énergie pour chaque segment voisé par rapport à l'énergie totale des segments voisés de la phrase.
- Débit: Il est défini par l'inverse de la moyenne des durées des segments voisés.
- Qualité vocale: On utilise les fréquences et largeur de bande des trois premiers formants. De plus on calcule les coefficients LPC de l'enveloppe spectrale.

Il existe une distinction assez fondamentale dans la direction de toutes les études sur les émotions. Certains proposent que les variations de la qualité vocale (ou du timbre hors contexte phonétique) soient parties intégrante de la prosodie. D'autres statuent que la qualité vocale et la prosodie sont indépendantes. Cette discussion découle directement d'un débat plus large sur la relation entre l'émotion et la cognition qui a été vivement animé dans les années 80 par Zajonc et Lazarus, les représentants des arguments pour la primauté de l'émotion et la primauté de la cognition. Zajonc (1980, 1982, 1984) propose l'indépendance partielle du système de l'émotion et du système cognitif, qui permet à l'émotion de se produire sans intervention de la cognition, tandis que

Lazarus (1982, 1984) insiste sur l'indispensabilité du processus cognitif dans la production de l'émotion, donc la nature postcognitive de l'émotion. (Chung, 2000)

Ainsi, certains mettent l'accent sur l'étude la prosodie (contours de f0...) comme (Murray, 1993), plutôt que sur la qualité vocale:

	f0	**Intensité**	**Débit**	**Qualité vocale**	**Autre**
Colère	- moyenne ++ - variabilité ++ - changements de hauteurs abruptes sur les accents	- intensité ++	- débit ++	- Soufflée	- articulation tendue
Peur	- moyenne ++ - variabilité ++	- intensité ++	- débit ++		- contours ascendants accentués - contours descendants atténués
Joie	- moyenne ++ - variabilité ++ - changements continus, inflexions vers le haut	- intensité ++	- débit ++ ou --	- soufflée, tonalité d'éclatement	- articulation normale
Tristesse	- moyenne -- - variabilité -- - inflexions vers le bas	- intensité --	- débit --	- résonante	
Dégoût	- moyenne -- - variabilité ++ - Inflexions fortes vers le bas	- intensité --	- débit ++	- Tonalité de coffre	- articulation normale

Émotions et paramètres de la voix: (Murray, 1993)

Remarque: Dans le tableau ci-dessus et dans tout le présent document:

° ++ veut dire: « grand »
° 00 veut dire: « moyen »
° -- veut dire: « petit »

Dans l'article de (de Mareüil *et al.*, 2000), les auteurs remarquent que le Français est une langue faiblement accentuée. Si l'ironie se traduit en Français par une moyenne de f0 élevée suivie d'une chute de f0 sur la dernière syllabe, l'étude de la qualité de la voix dans le cas des émotions: ennui, joie et colère froide (nous reviendrons plus tard sur ces deux dernières) s'avère plus pertinente que l'étude de la prosodie. Leurs résultats sont ici reportés:

	f0	**Intensité**	**Débit**	**Qualité vocale**	**Autre**
Colère	- moyenne ++ - variabilité ++	- intensité ++	- débit ++		- débit, énergie, f0 et contours de f0 sont accrus durant la phrase.
Peur	- moyenne ++ - variabilité ++ - f0 initial = max (f0)	- intensité ++	- débit ++		- contours ascendants accentués - contours descendants atténués
Surprise	- contour descendant sur la dernière syllabe		- début lent		- Emphase sur la pénultième syllabe
Joie	- moyenne ++ - variabilité ++	- intensité ++	- débit ++		
Tristesse	- moyenne -- - variabilité --	- intensité --	- débit --		
Dégoût	- moyenne -- - les contours sont inversés - f0 final = min (f0)	-	- débit --		

(De Mareüil et al., 2000)

Une étude intéressante et originale est celle de (Drioli *et al.*, 2003) qui met en parallèle la tension musicale (dissonance) aux émotions négatives:

	f0	**Intensité**	**Débit**	**Qualité vocale**	**Autre**
Colère	- Moyenne 00 - Variabilité ++	- Intensité ++	- débit ++	- qualité « dure » - prédominance des voyelles accentuées	- intervalle d'un triton (7/5) par rapport au neutre
Peur	- Moyenne ++ - Variabilité --	- Intensité 00	- débit 00	- qualité « soufflée »	- Intervalle d'une seconde majeure par rapport au neutre
Dégoût	- Moyenne 00 - Variabilité ++	- Intensité 00	- débit --	- qualité « grinçante » adoucie sur les voyelles accentuées	- Intervalle d'une seconde majeure par rapport au neutre
Joie et surprise	- Moyenne ++ - Variabilité ++ (seulement sur les voyelles accentuées)	- Intensité 0+	- débit ++	- qualité « claire » - voix soufflée sur les attaques de voyelles accentuées - voix dure sur les attaques de voyelles accentuées	- Les deux sont difficiles à distinguer - Intervalle d'une octave augmentée par rapport au neutre
Tristesse	- Moyenne ++	- Intensité 00	- débit 00	- qualité « sombre »	

(Drioli et al., 2003)

Les auteurs remarquent que la colère, le dégoût, la peur et la tristesse présentent d'évidents intervalles dissonant en ce qui concerne le rapport de leurs fréquences fondamentales moyennes à celle du neutre. De plus, ils présentent quatre descripteurs de la qualité vocale non répertoriés jusqu'ici:

° Rapport signal/bruit (Harmonique/bruit)

° Drop-off energy = pente spectrale au delà de 1000 Hz (estimée par moindres carrés)

° Rapport d'énergie au-delà de 1000Hz / Énergie en-dessous de 1000Hz

° Spectral flatness measure: rapport entre la moyenne arithmétique et la moyenne géométrique de la distribution spectrale d'énergie.

Ainsi, il semble que la qualité vocale soit déterminante dans la reconnaissance (et donc dans la synthèse) des émotions. Cependant, comme en musique, le timbre est difficile à définir devant sa variabilité. Dans (Abadjieva *et al.*, 1993), Il n'est définit que par une distinction spectrale: énergie dans les hautes ou les basses fréquences:

	Domaine fréquentiel	Domaine temporel	Qualité de la voix
Joie	- f0 moyen élevé - variation du f0 dynamique	- débit rapide - structure rythmique avec accentuation régulière	- intensité élevée mais pas autant que pour la colère - voix légèrement aspirée
Colère	- f0 moyen élevé ou modéré - variation du f0 dynamique - contour de f0 fortement descendant en fin de phrase	- débit plus rapide que la voix neutre mais moins que pour la joie	- intensité élevée - voix aspirée et tendue - grande énergie dans les hautes fréquences
Tristesse	- f0 moyen au niveau de la voix neutre - peu de variation du f0	- débit lent - rythme avec des pauses régulières	- intensité basse sans variation - articulation moins précise
Peur	- f0 moyen légèrement élevé - grande variation du f0, mais pas autant que pour la joie ou la colère	- débit plus lent que pour la joie et la colère mais plus rapide que la voix neutre - pauses irrégulières	- intensité basse - articulation précise - peu d'énergie dans les basses fréquences

Les indices acoustiques des émotions primaires, (Abadjieva et al., 1993)

(Laukka, 2004), dans sa thèse, décrit la qualité de vocale de manière plus approfondie en donnant la position et la largeur du premier formant ainsi que l'allure de la forme d'onde glottique. Dans l'hypothèse source filtre, classique dans l'étude de la voix, la forme d'onde glottique est riche en renseignement car elle contient beaucoup d'informations d'ordre physiologique sur les émotions (tension/détente).

Malheureusement, son extraction est ardue car il s'agit d'un problème inverse complexe. Cependant, on peut sous certaines hypothèses, déduire du signal une allure de la forme d'onde glottique et donc, s'en servir pour la description des émotions.

	f0	Intensité	Débit	Qualité vocale	Autre
Colère	- Moyenne ++ - Variabilité ++ - Maximum ++ - Contours montants - Jitter ++	- Intensité ++ - Variabilité ++	- débit ++ - pause -- - irrégularités micro structurelles	- énergie des Fréquences élevées ++ - F1 (premier formant): ° fréquence ++ ° largeur de bande -- - Forme d'onde glottique raide	- articulation précise
Peur	- Moyenne ++ - Variabilité -- - Contours montants	- Intensité – (sauf en cas de panic) - variabilité ++	- débit ++ - irrégularités micro structurelles	- énergie des fréquences élevées -- - F1 (premier formant): ° fréquence -- ° largeur de bande ++ - Forme d'onde glottique arrondie	
Dégoût	- Moyenne -- - Variabilité 00 - Contours descendants	- Intensité 00 - variabilité 00	- débit -- - pauses ++	- énergie des fréquences élevées 00 - F1 (premier formant): ° fréquence -- ° largeur de bande ++	- articulation précise - voice onsets lents
Joie	- Moyenne ++ - Variabilité ++ - Maximum ++ - Contours montants	- variabilité ++	- débit ++ - irrégularités micro structurelles	- énergie des fréquences élevées 00 - F1 (premier formant): ° fréquence ++ ° largeur de bande ++ - Forme d'onde glottique raide	- voice onsets rapides
Tristesse	- Moyenne -- - Variabilité -- - Maximum -- - Contours descendants - jitter --	- Intensité -- - variabilité --	- débit -- - pauses ++ - irrégularités micro structurelles	- énergie des fréquences élevées -- - F1 (premier formant): ° fréquence -- ° largeur de bande ++ - Forme d'onde glottique arrondie	- articulation dégagée
tendresse	- Moyenne -- - Variabilité -- - Contours descendants	- Intensité --	- débit -- - irrégularités micro structurelles	- énergie des fréquences élevées --	- voice onsets lents

Résultats de la thèse de P. Laukka (Laukka, 2004)

M. Schröder est l'administrateur du site HUMAINE qui permet de centraliser de nombreuses études sur les émotions (faciales, vocales...). Dans un article (Schröder, 2001), il donne des exemples de règles prosodiques efficaces pour l'expression d'émotions dans la parole synthétique.

Émotion **Étude** **Langue** **Taux de reconnaissance**	**f0**	**Intensité**	**Débit**	**Qualité vocale**	**Autre**
Colère (Murray. Arnott, 1995) Anglais	- moyenne: + 10% - Intervalle: 9 semitons	- intensité: +6dB	- débit: +30 mots par min	- Laryngualisation: +78% - Fréquence de F4: -175 Hz	- f0 élevé sur les voyelles stressées: Emphatique: +40% Primaire: +20% Secondaire: +10%
Peur (Burkhardt Sendlmeier, 2000) Allemand 52%	- moyenne: +150% - Intervalle: +20%		- débit: +30%	- voix de fausset	
Surprise (Cahn, 1989) Américain 44%	- moyenne: -8 - intervalle: +8 - Contours raides	- intensité: +5	- débit: +4 - pauses --	- Brillance: -3	
Joie (Burkhardt Sendlmeier, 2000) Allemand 81%	- moyenne: + 50% - Intervalle: +100%	- intensité: +6dB	- débit: +30%	- Propagation au niveau des lèvres: => F1/F2: +10%	- f0 élevé sur les voyelles stressées: Emphatiques: +100% Syllabes entre: -20%
Ennui (Mozziconacci, 1998) Hollandais 94%	- moyenne basse en fin de phrase - Intervalle: 4 semitons	- intensité --	- débit: -150%		- Patterns intonatifs particuliers
Tristesse (Cahn, 1989) Américain 91%	- moyenne: -5 - intervalle: -5	- intensité: -5	- débit: -10 - pauses +10	- Brillance: -9 - souffle: +10	- précision de l'articulation: -5

(Schröder, 2001)

Les algorithmes de classification entraînés sur un corpus lors d'un apprentissage supervisé permettent de déduire la pertinence des descripteurs et ainsi de les classer. (Schuller *et al.*, 2005) classe ces descripteurs grâce à une sélection réalisée par machine à vecteur de support (SVM-SFFS). Au final, il ne retient que les suivants, classés dans l'ordre de préférence:

1. Gradient maximal de f0
2. Moyenne de f0
3. Moyenne de l'énergie normalisée
4. Moyenne du gradient de f0
5. Taux de passage par zéro du signal.
6. Variance de la hauteur
7. Maximum relatif de f0
8. Moyenne des durées des silences
9. Gradient maximal d'énergie
10.Intervalle de f0
11.Distance entre les moyennes de f0 au début et à la fin de la phrase
12.Variance des durées des sons voisés
13.Valeur médiane de l'énergie sur le temps de montée
14.Valeur médiane des durées des silences
15.Moyenne des durées des sons voisés
16.Énergie spectrale en dessous de 250 Hz
17.Distance entre les variances de l'énergie des points initial et final de la phrase.
18.Énergie moyenne sur le temps de chute.
19.Distance entre les moyennes de l'énergie des points initial et final de la phrase.
20.Maximum relatif d'énergie

Classement de descripteurs issu d'un apprentissage supervisé (Schuller et al., 2005)

Cependant, ces résultats sont à prendre en douceur car ils dépendent très fortement du corpus étudié, du type d'émotions analysées... En règle générale, il n'y en a pas dans l'étude des émotions car celles-ci sont toujours dépendantes de la perception d'un ou de plusieurs évaluateurs (lors de test perceptifs...). Nous souhaitons simplement arriver à décrire pour un acteur, quelques stratégies de l'expression émotionnelle. Comme celle-ci est dirigée, elle se traduit par des faits prosodiques conventionnels et donc, connus de tous (partie cognitive de l'émotion).

Il est à noter que s'il n'existe pas de consensus au niveau des émotions étudiées, c'est pour un double raison:

- La première provient directement du fait qu'il n'existe pas de consensus dans la définition même des émotions.

- La seconde relève des différentes motivations de chacun, notamment en vue d'applications largement éloignées.

La complexité est accrue lorsqu'on se rend compte que même entre catégories désignées à l'avance, il peut y avoir des interférences au niveau de leurs perceptions, comme le montre les exemples de matrices de confusion suivantes:

	Calme	Colère	Tristesse	Confort	Joie
Calme	**76**	3	4	14	3
Colère	0	**92**	0	0	8
Tristesse	8	0	**76**	16	0
Confort	15	0	5	**77**	3
Joie	4	20	0	8	**68**

Matrice de confusion (Oudeyer, 2003)

	Triste	Ennui	Joie	Colère	Neutre
Triste	**0.42**	0.3	0.16	0.04	0.09
Ennui	0.28	**0.36**	0.18	0.05	0.13
Joie	0.15	0.12	**0.44**	0.17	0.12
Colère	0.02	0.8	0.19	**0.7**	0.01
Neutre	0.16	0.25	0.2	0.02	**0.37**

Matrice de confusion (Yacoub, 2003)

La confusion la plus largement citée dans la littérature est assez paradoxale: « Les paramètres acoustiques peinent à discriminer les expressions de joie exaltées et les confondent avec les émotions fortement activées telle que la colère froide » (Bänziger, 2004).(Chung, 2000) présentent des résultats mettant en emphase la confusion entre joie et colère. Les valeurs en gras sont les plus significatives.

Paramètres acoustiques	**Catégories d'émotions**		
	Joie	Neutre	Colère
Fo moyen (Hz)	**249,75**	222,07	229,81
Fo maximum (Hz)	**348,94**	302,60	**347,17**
Fo minimum (Hz)	124,69	109,71	**76,55**
Fo Moy 20% Bas (Hz)	186,69	169,02	**127,83**
Plage de Fo (Hz)	224,25	192,88	**270,62**
Jitter (%)	0,60	0,80	0,90
Shimmer (%)	0,03	0,03	0,07

Paramètres acoustiques	Catégories d'émotions		
	Joie	Neutre	Colère
Débit de parole (N. de syll./sec)	7,00	6,40	6,30

(Chung, 2000)

Remarques:
Définition du Jitter: Dans cette analyse, le jitter est mesuré par l'estimation du pourcentage de la différence des valeurs de f0 entre les cycles adjacents, dont le calcul mathématique peut être exprimé comme suit :

$$\text{jitter } (\%) = \frac{100\,N \sum_{1}^{N-1} |F0_{i+1} - F0_i|}{(N-1) \sum_{1}^{N} F0_i}$$

Définition du Shimmer: Dans cette analyse, le shimmer est mesuré par l'estimation du pourcentage de la différence des valeurs d'amplitude entre les cycles adjacents, dont le calcul mathématique peut être exprimé comme suit :

$$\text{shimmer } (\%) = \frac{100\,N \sum_{1}^{N-1} |A_{i+1} - A_i|}{(N-1) \sum_{1}^{N} A_i}$$

Il est à noter que ces deux derniers indices ne paraissent pas pertinents à la vue des résultats de (Chung, 2000). Plusieurs discussions sur ces descripteurs mettent en avant de nombreuses précautions à prendre pour les utiliser, notamment en ce qui concerne la taille de la fenêtre d'observation (paramétré ici par N).

Si notre perception ne nous fait pas défaut face à cette distinction, de nombreux algorithmes de classification s'y trompent car les expressions de ces deux émotions antagonistes d'un point de vue ontologique, possèdent généralement une moyenne de fréquence fondamentale élevée. Ainsi, certaines catégories émotionnelles sont moins bien discriminées que d'autres.

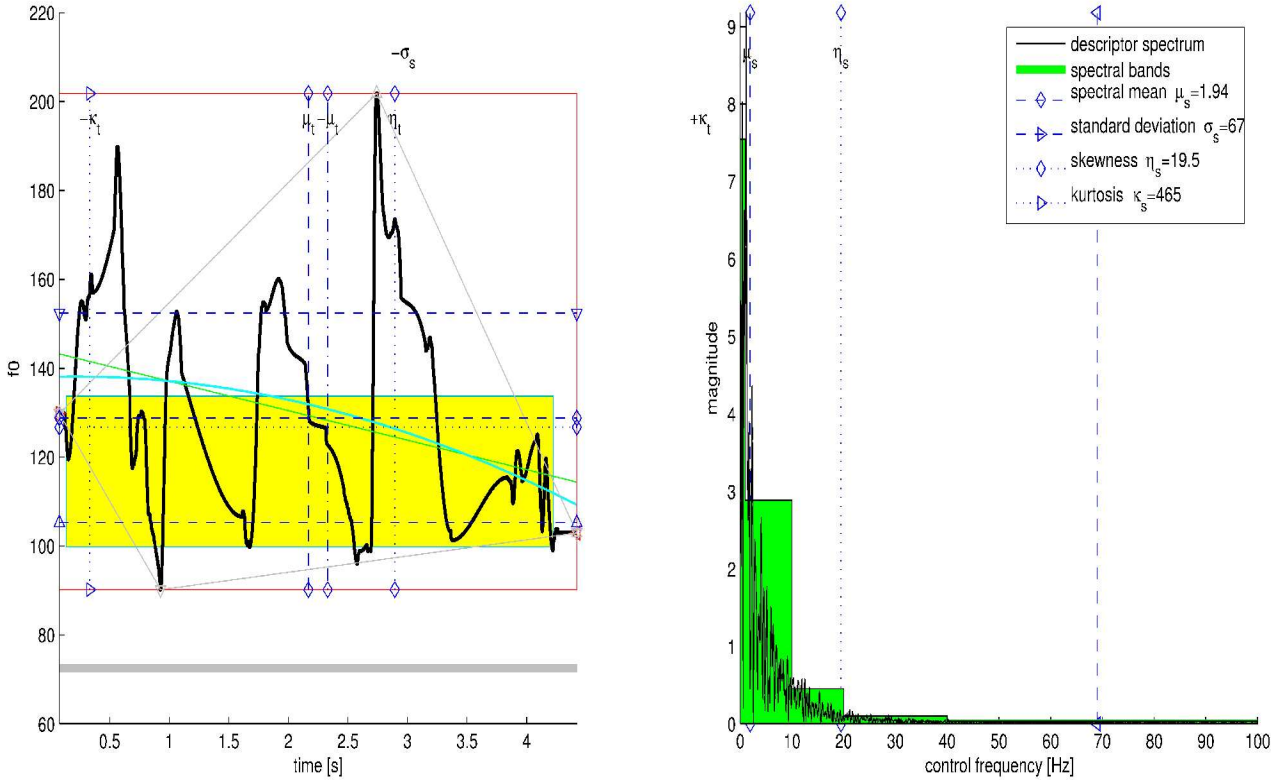
Nous émettons donc une hypothèse: Si notre perception des émotions est catégorielle, il ne peut exister de descripteur acoustique les discriminant, mais seulement des sous-ensembles de descripteurs dont les compositions sont propres aux affects et donc, différentes d'une émotion à l'autre. Cela veut dire qu'à chaque catégorie émotionnelle correspond un sous-ensemble de descripteurs acoustiques pertinents pour la décrire et que, par conséquent, toute catégorisation à partir d'un même ensemble est vaine.

IV.2.Descripteurs utilisés:

Les évolutions dynamiques des courbes de fréquence fondamentale, de débit et d'énergie décrite dans la partie II de ce rapport, sont modélisées temporellement par des valeurs caractéristiques. Pour chacune des unité segmentales (phone, diphone, semiphone, syllabe) et supra-segmentales (groupes prosodiques, phrases), nous calculons sur leur horizon temporel:

- des descripteurs globaux:
  - la moyenne arithmétique, la moyenne géométrique, la variance
  - la valeur de début (valeur médiane sur un petit horizon de 10 ms au début de l'unité)
  - la valeur de fin (valeur médiane sur une petit horizon de 10 ms à la fin de l'unité)
  - le minimum, le maximum et l'écart absolu entre ces deux valeurs
- des descripteurs modélisant l'évolution temporelle:
  - l'interpolation linéaire de la courbe (« slope »)
  - l'interpolation cubique par un polynôme de Legendre d'ordre 2 et l'énergie du résiduel (pour juger de la vraisemblance de cette estimation)
  - les centres de gravité et d'antigravité temporel (qui définit la location de la plus importante dépression ou élévation de la courbe)
- des descripteurs modélisant le spectre:
  - Les 4 premiers moments centrés du spectre du descripteur (moyenne, variance, l'obliquité (« skewness ») et « kurtosis »)
  - L'énergie contenue dans 5 bandes de fréquences (0, 1, 10, 20, 40, 100 Hz)

Ces valeurs caractéristiques calculées automatiquement lors de l'importation d'un fichiers de données (ex: fréquence fondamentale) sont tracées ci-dessous pour une unité de type phrase d'une durée d'environ 4 secondes. (voir (Schwarz, 2000, 2003a, 2003b, 2004) et (Beller et al., 2005))



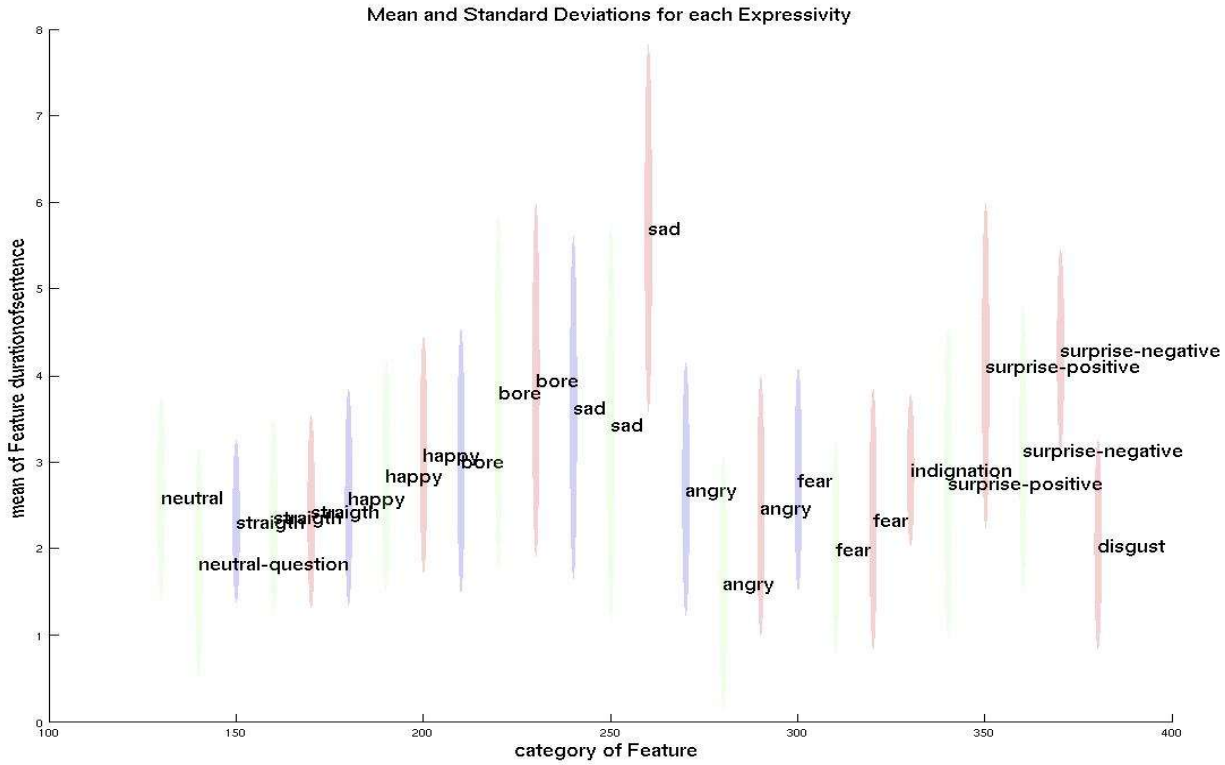
V.Résultats et discussions:

V.1.Présentation des Résultats:

Nous allons tenter d'illustrer et de valider l'hypothèse citée précédemment. Si notre perception des émotions est catégorielle, cela veut dire qu'un petit indice (respiration forte, hauteur anormalement haute sur un phonème...) peut faire basculer une phrase neutre en une phrase affectée pour notre écoute. Cela veut dire que nous ne catégorisons pas les émotions par l'examen simultané de l'évolution de nombreuses variables acoustiques, mais plutôt par des indices locaux et souvent différents selon les émotions exprimées. D'où l'hypothèse que tous les affects ne peuvent être décrits par le même ensemble de descripteurs, mais chacun, par un sous-ensemble pertinent. Nous allons tenter de montrer par quelques observations de la base de données l'importance d'une telle hypothèse.

Voici, ci-dessous, l'examen des données correspondant à la requête suivante: Pour toutes les catégories émotionnelles (abscisse), tracer la valeur moyenne (ordonnée) et l'écart type (largeur verticale) des durées (en secondes) des phrases prononcées avec l'affect correspondant. Le degré d'activation de l'affect (faible, moyen et fort) est représenté par la couleur de la tâche (respectivement bleu, vert et rouge).

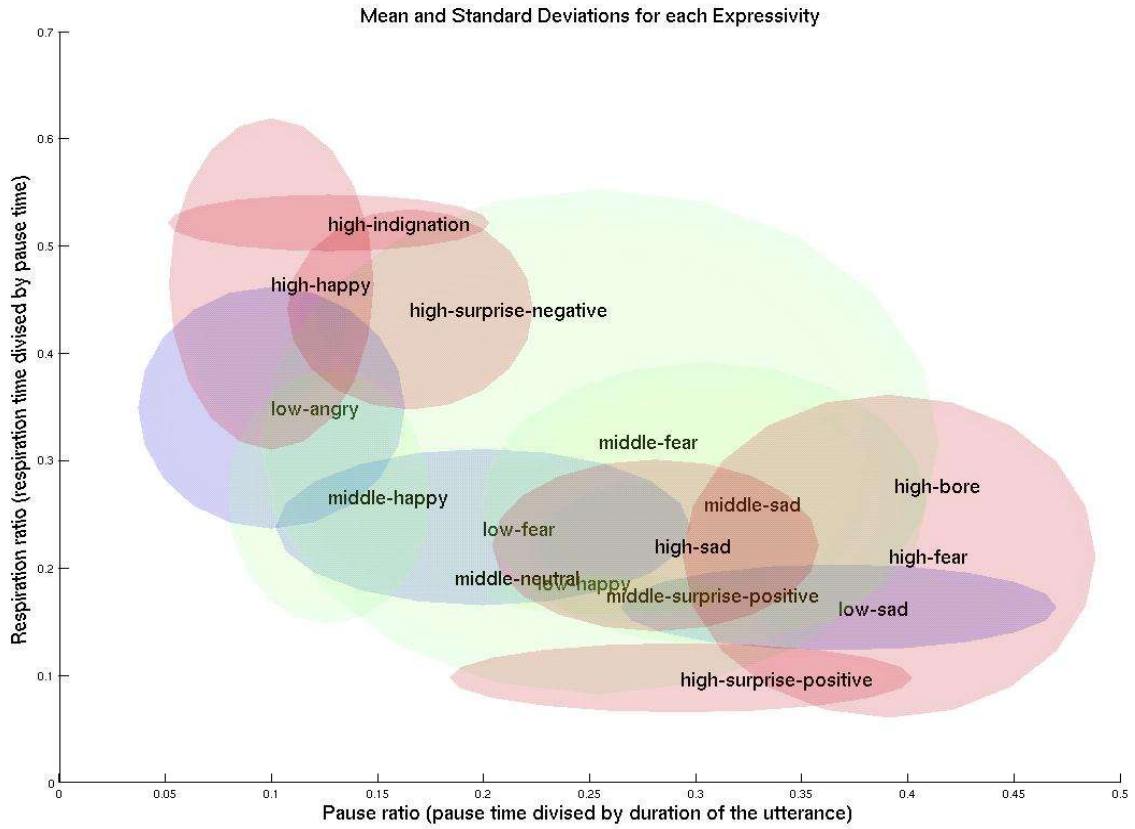
Dans cet exemple, on voit clairement que la durée des phrases se révèle un facteur déterminant pour la tristesse forte (« sad » en rouge). Ceci est dû à la présence de nombreuses pauses, parfois sonores par la présence de respirations. En revanche, ce « descripteur » n'est pertinent que pour cette catégorie émotionnelle et ne permet pas de discriminer la joie de la colère.



Dans la suite, nous présenterons l'essentiel de nos résultats avec ce genre de graphiques.

V.2.Pauses :

Les pauses sont un facteur très important dans la prosodie (Zellner, 1998). On peut étudier leur influence en examinant le rapport entre le temps de pause et le temps de parole (en abscisse). Pour plus de précision, nous avons décidé d'étudier aussi le rapport entre le temps de respiration (quand celles-ci sont sonores) et le temps de pause (durant laquelle a lieu la respiration) (en ordonnée):



On observe que pour la peur moyenne (« middle fear »), la variance des résultats est trop forte pour en tirer une quelconque information. En revanche, la forte indignation (« high indignation ») et la surprise positive présente deux comportement bien différents et marqués (variance faible) vis à vis du rapport entre le temps de respiration et le temps de pause. Ainsi, le rapport élevé de l'indignation forte révèle que les pauses courtes sont majoritairement sonores (tout comme pour la joie forte). Inversement, elles sont plutôt longues et silencieuses pour la forte surprise positive et la peur.

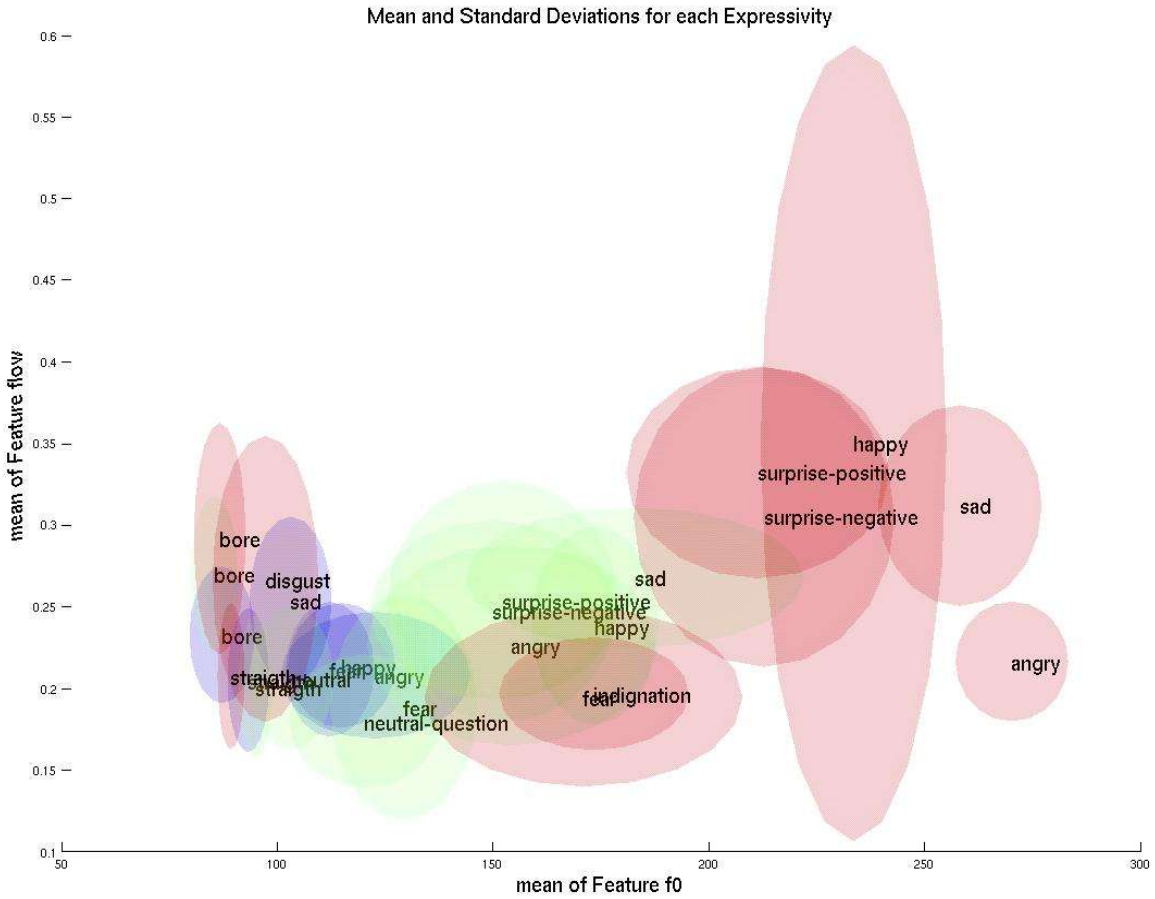
V.3.espaces discriminants:

On voit apparaître sur ce dernier graphique, la notion d'espace discriminant. En effet, en traçant une classification sur une échelle bidimensionnelle (deux descripteurs), il apparaît une distance plus grande entre la forte surprise négative et la forte surprise positive que sur une seule dimension.

La poursuite des travaux lors d'une thèse ira dans ce sens et permettra de mettre en oeuvre des algorithmes de classification (machine à vecteur de support, K plus proches voisins...). Ils permettront de déterminer les meilleurs espaces discriminants qui maximisent les distances entre

affects. Ainsi, grâce à une phase d'apprentissage, on parviendra non seulement à une classification automatique des affects, mais surtout à une classification automatique des descripteurs qui permettent le mieux de discriminer ces affects.

Pour le moment, il est déjà très intéressant d'effectuer quelques mesures « intuitives » sur les données. L'espace discriminant qui suit, parait être le plus pertinent dans la littérature: En abscisse, on représente la moyenne et la variance de la fréquence fondamentale et en ordonnée, la moyenne et la variance du débit syllabique:



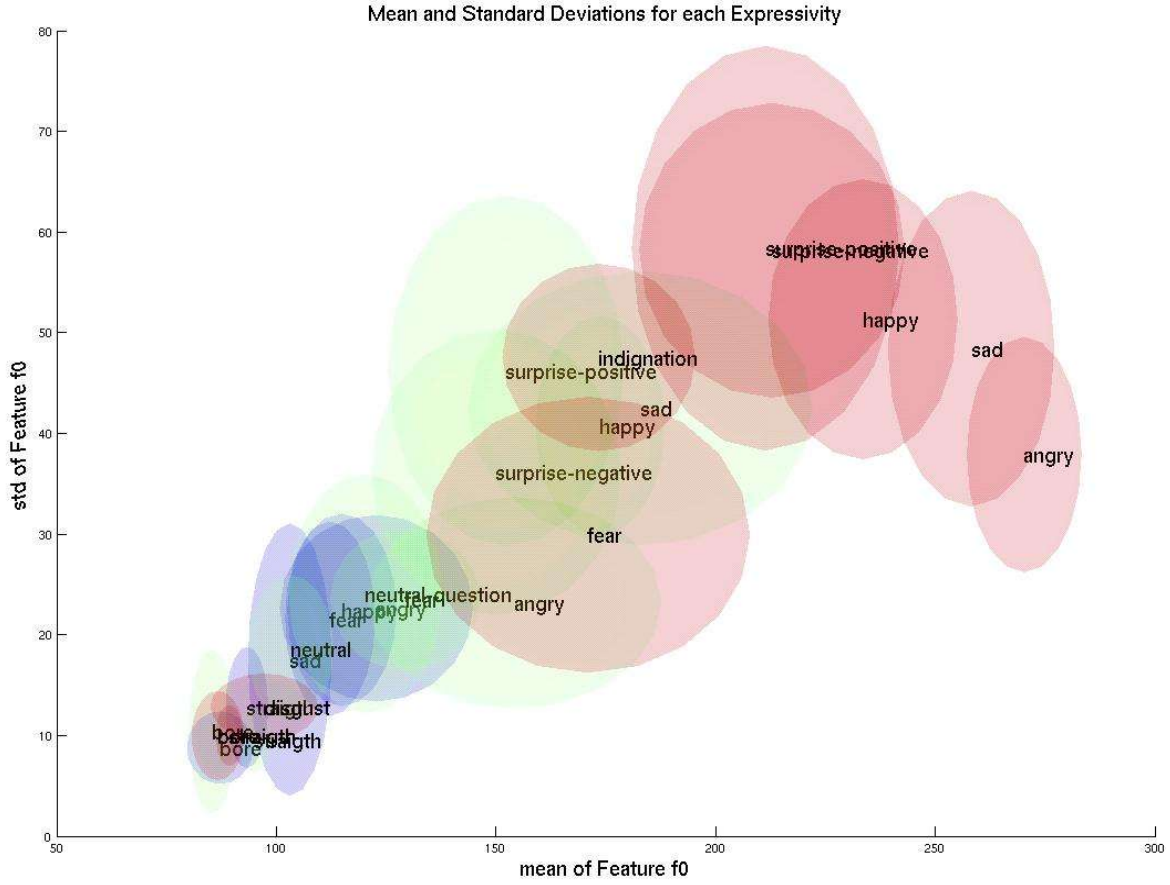
La première remarque vient de la croissance de la moyenne de f0 en fonction de la force de l'activation de l'affect. Les zones rouges à droites montrent que plus l'affect est exprimé avec force, plus la fréquence fondamentale croit (sauf dans les cas de l'ennui et du dégoût pour lesquels la variation est inverse). Cette tendance se retrouve partout dans la littérature. Cet indice seul ne permet pas de différencier la joie de la colère (Bänziger, 2004), (Chung, 2000).

Si l'on prend maintenant en compte le débit, on peut distinguer la tristesse de la colère puisque celle-ci s'exprime par un débit plus rapide. La variance large du débit syllabique de la joie forte montre que celle-ci est exprimée par des variations fortes du rythme. Au contraire, il semble que la colère s'exprime par un débit plus rapide et recto rythmique.

Examinons plus en détail ces deux paramètres fondamentaux de la prosodie ainsi que l'intensité et le timbre vocal.

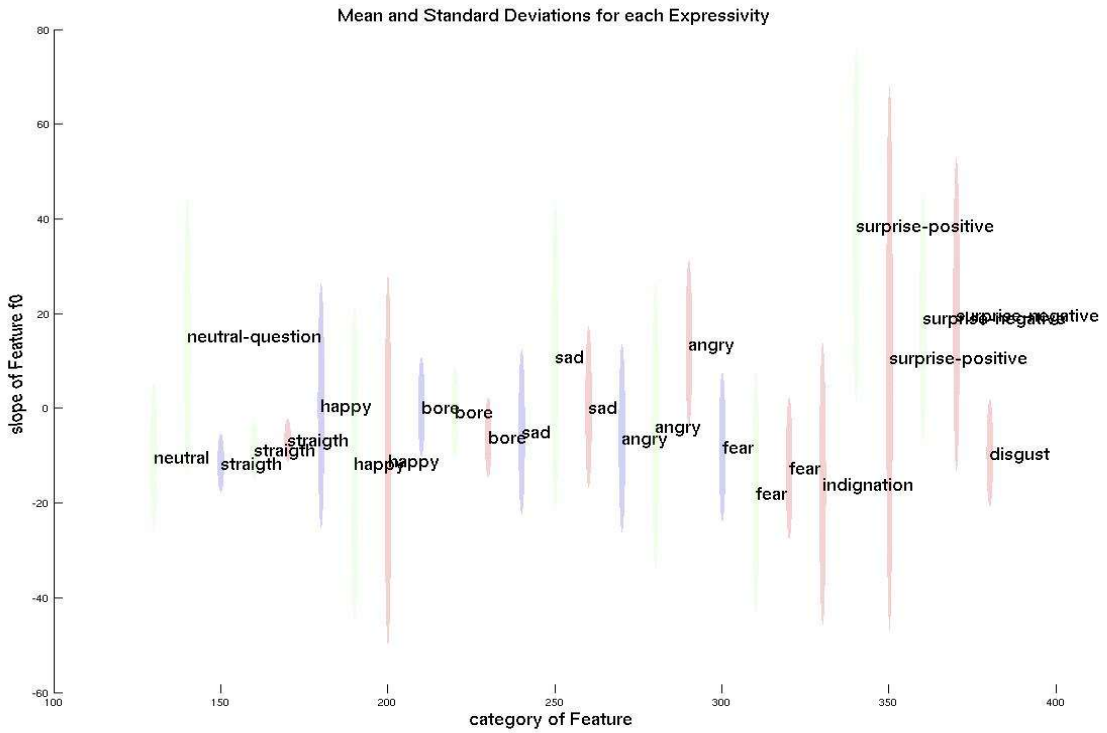
V.4.Fréquence fondamentale f0:

Nous parlons du caractère recto rythmique pour la colère qui se manifeste aussi par son caractère recto tonal. En effet, si l'on observe la moyenne (en ordonnée) des variances de f0 au cours d'une phrase, on voit que la colère, la peur, l'ennui et le dégoût se distingue par leur rectitude au niveau de la hauteur. L'inverse se produit pour la joie et la surprise (négative ou positive) qui montrent une grande variabilité de hauteur au sein d'une phrase.



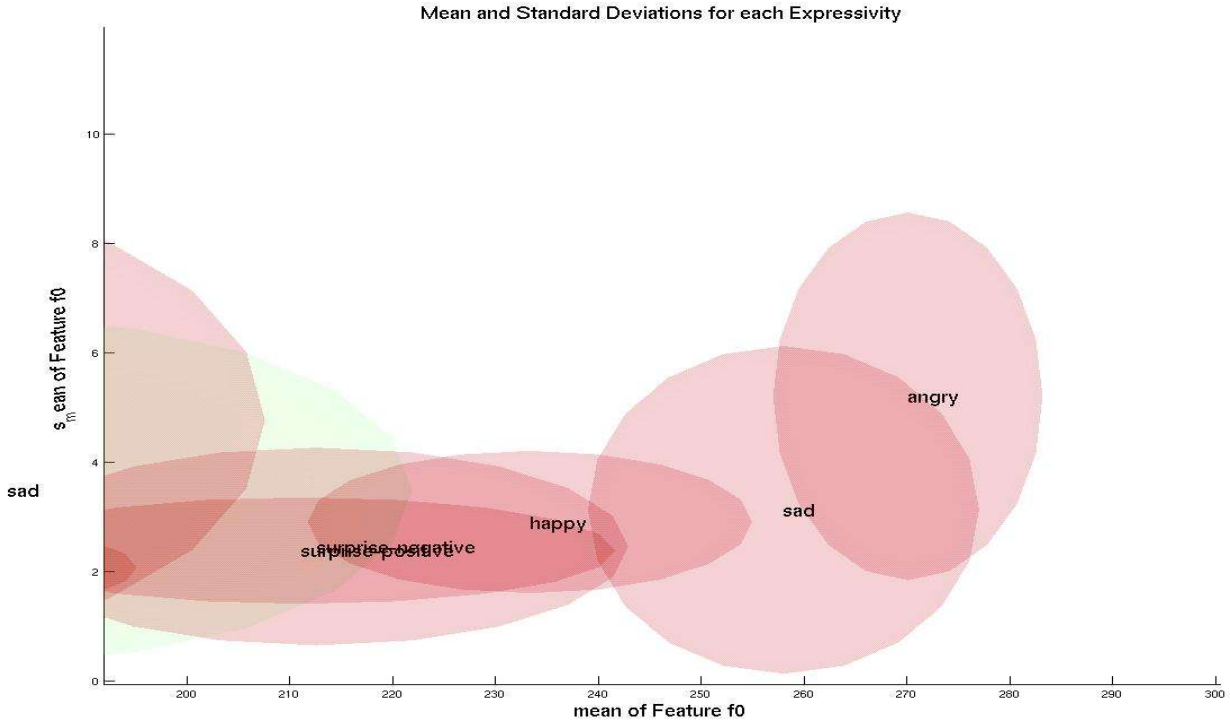
Cette grande variabilité et aussi due à la modalité de la phrase (question/affirmation), comme le confirme la figure suivante représentant la « direction » globale des phrases. Elle présente pour chaque affect, la moyenne des valeurs caractéristiques correspondant à l'interpolation linéaire (« slope ») de la fréquence fondamentale sur l'horizon d'une phrase. Lorsque cette moyenne est positive, cela correspond à un ton montant très souvent associé au ton interrogatif. Aussi il n'est pas étonnant de voir que la surprise et la question neutre domine dans ce domaine. Toutefois il existe des questions posées avec un ton descendant; Sans généraliser, la figure suivante nous donne la tendance générale de la stratégie menée par l'acteur pour exprimer la modalité interrogative.

Les autres affects possèdent pour la plupart une moyenne légèrement négative qui traduit la déclinaison naturelle des phrases possédant une origine purement acoustique: En fin de souffle à la fin d'une phrase, la pression sur les cordes vocales diminuent et la hauteur chute. Cette déclinaison a toutefois été amplifiée lors des expressions de l'indignation et du dégoût.



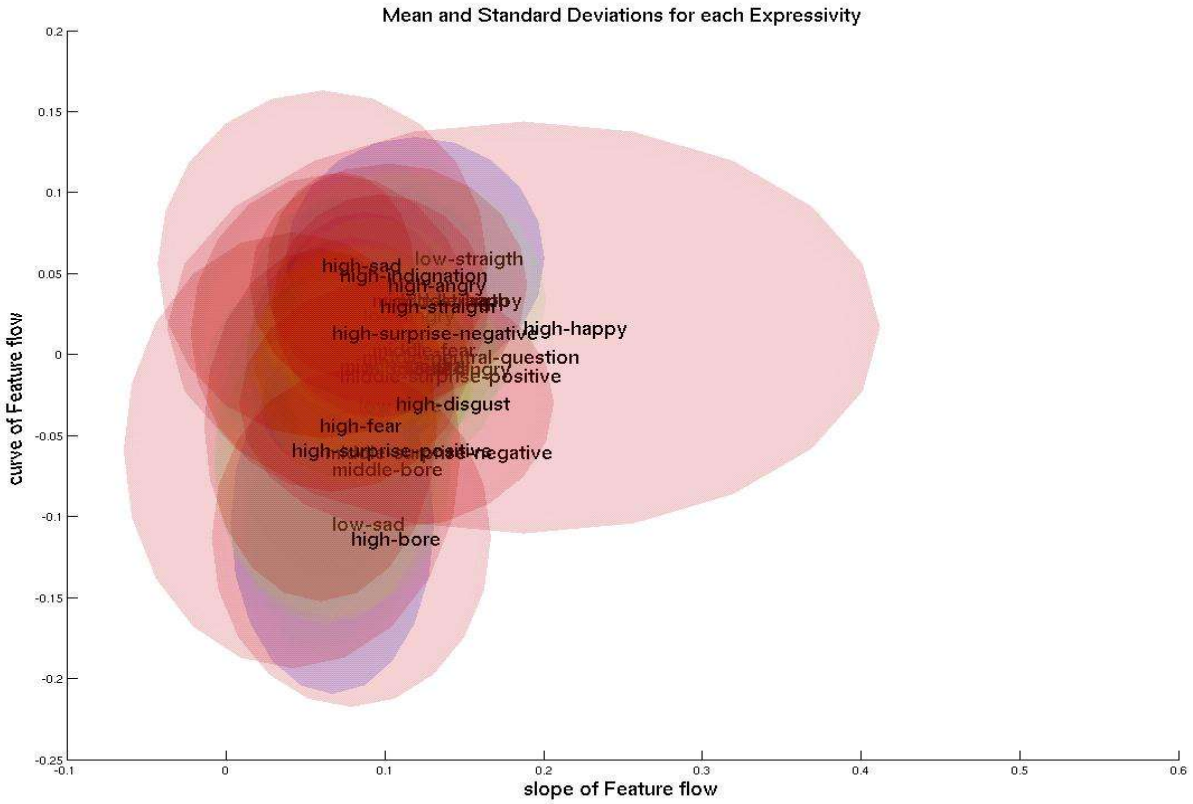
Jitter:

La perturbation de la fréquence fondamentale est un indice important (voir Partie IV.A). De manière à évaluer son importance, nous traçons (en ordonnée) la moyenne fréquentielle du spectre en fonction de la moyenne de la fréquence fondamentale (en abscisse). Ce genre de perturbation devient visible pour un fort degré d'activation, c'est pourquoi nous ne présentons ces valeurs que pour les états fortement activés. La croissance de la moyenne du spectre de la fréquence fondamentale signifie la présence de hautes fréquences dans la hauteur dues aux perturbations d'origine physiologique. La forte présence de Jitter dans la colère correspond à la littérature.



V.5.Débit de parole:

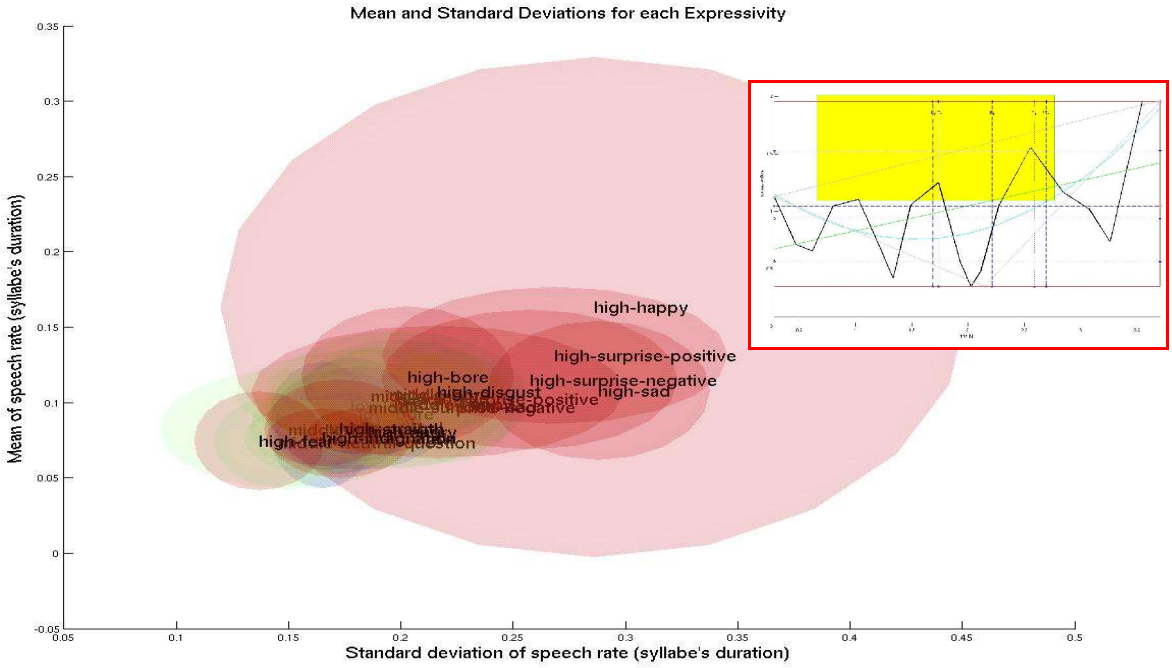
Nous avons vu qu'une forte variance du débit syllabique était caractéristique de la joie. Il existe un autre indice pertinent concernant la caractérisation de cet affect dans l'évolution temporelle de son débit. Très souvent pour exprimer la joie, l'acteur finit les fins de phrases en accentuant la décélération finale. Ce phénomène a d'ailleurs lieu dans la plupart des phrases, quelque soit l'affect (« slope » > 0)), mais il est accentué dans le cas de la joie comme le montre la figure ci-dessous.



Sur cette figure, on peut aussi observer qu'il existe toute sorte de « courbure » du débit. C'est à dire que le contour moyen de débit est variable et ne semble pas au premier abord déterminant; Bien que l'ennui fort (mean(curve) ~= -O,1) présente une courbure négative assez forte qui indique que l'acteur a souvent ralenti en milieu de phrase avant d'accélérer pour exprimer l'ennui.

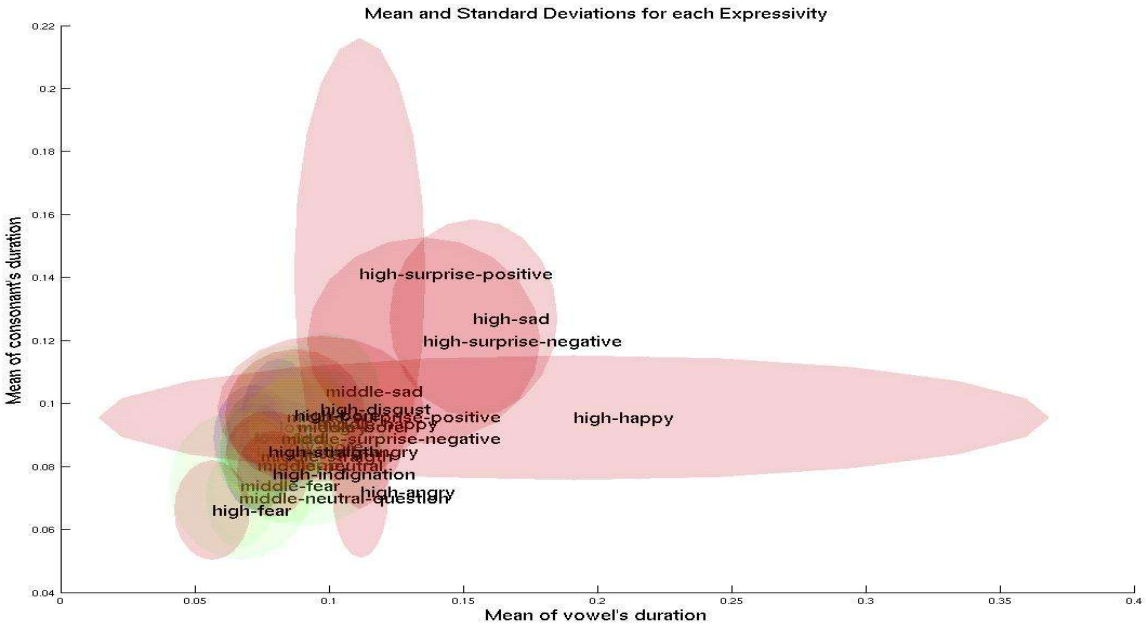
La joie semble aussi se démarquer des autres par la grande variabilité de son débit, comme l'illustre la courbe suivante. Non seulement elle présente le débit moyen le plus lent, mais elle est exprimée par une très grande et significative variabilité du rythme syllabique. Ce phénomène est surtout dû à ce que la joie a été exprimée par l'acteur par des voyelles tenues, presque chantées.

Un exemple d'évolution du débit syllabique sur une phrase hautement joyeuse (Phrase 170 du corpus) est présenté dans l'encadré. Si le débit syllabique possède une variance forte, cela ne veut pas dire que la phrase devient arythmique. Au contraire, il se dégage de cette variabilité, des zones plus claires de mises en emphase temporelle qui conduit naturellement à la perception d'un rythme musical. Une étude précise des occurrences relatives de ces mises en emphase dans le domaine temporelle, nous permettrait de mieux appréhender ce phénomène.

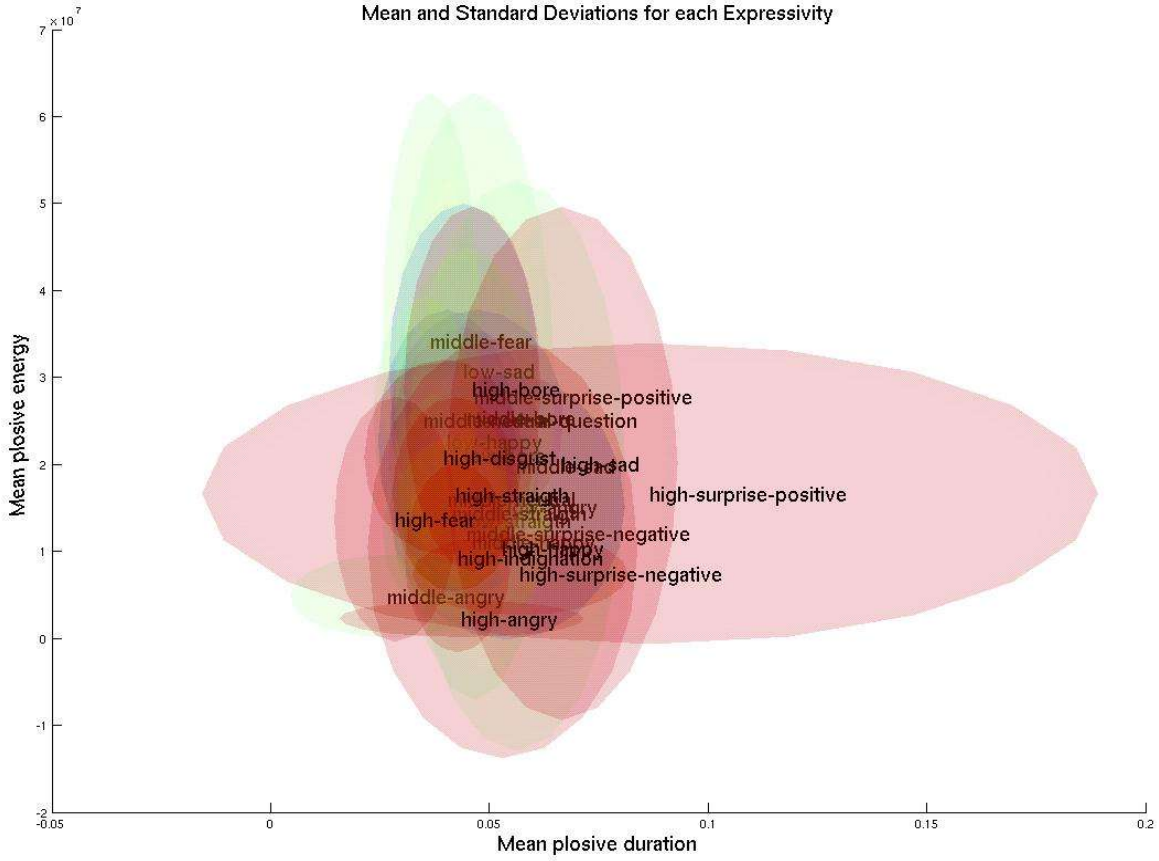


L'importance de l'aspect rythmique dans l'expression de la joie est un indice majeur pour la distinguer de la colère. Comme cité dans la littérature, ces deux affects se situent dans la même tessiture. Seul la différence de timbre a été soulignée jusqu'à présent. Hors on voit apparaître ici une autre différence majeure en ce qui concerne leurs débits et leurs variations.

Pour mettre en lumière ce fait, nous avons transformer des phrases joyeuses en phrases colériques en raccourcissant simplement les voyelles et en rallongeant les consonnes. Le résultat sonore est sans équivoque et sera présenté lors de la soutenance. Cette expérience souligne l'importance du rapport entre la durée des voyelles et celle des consonnes. C'est pourquoi nous traçons dans la figure suivante la moyenne des durées des voyelles (la largeur horizontale correspond à la variance de ces moyennes calculées chacune pour une phrase), en abscisse et la moyenne des durées des consonnes (la largeur verticale correspond à la variance de ces moyennes calculées chacune pour une phrase), en ordonnée.



On peut voir apparaître dans le cas de la joie, une plus grande proportion de voyelle que de consonnes (éloignement de la 1ère bissectrice). Cela est du au fait que des voyelles sont énormément allongées (chantées) comme l'exprime la taille de la variance. On remarque aussi sur cette figure la « tâche » correspondante à la forte surprise positive. Celle-ci se décline par le fait qu'elle est la seule à se trouver au-dessus de la 1ère bissectrice. La variance verticale large nous amène à penser que certaine consonnes sont très allongées, s'il on compare à notre interprétation dans le cas de la joie. Ce fait est confirmé par l'écoute des phrases qui présentent une accentuation forte de la durée des plosives (« p », « t », « k », « b », « d », « g ») ou plus précisément de leurs temps avant explosion.

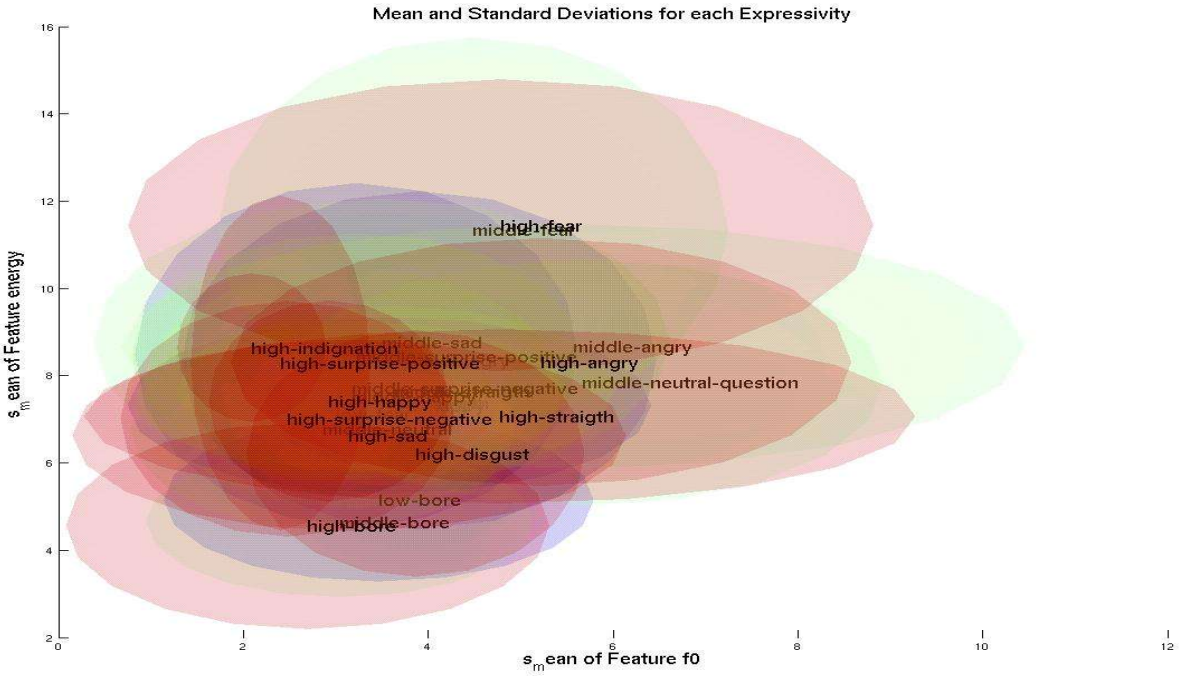


Dans ce stage, nous avons tenter de mettre l'accent (si je puis dire) sur l'aspect temporel de la parole parce que nous avons créé un descripteur dynamique du débit grâce à l'hypothèse d'une fréquence syllabique constante. Jusqu'ici, nous n'avons rencontré dans la littérature que des descripteurs globaux sur le temps d'une phrase ou d'un énoncé. Hors, on voit l'importance pour la qualification de la joie de ce descripteur dynamique.

V.6.Énergie:

Shimmer:

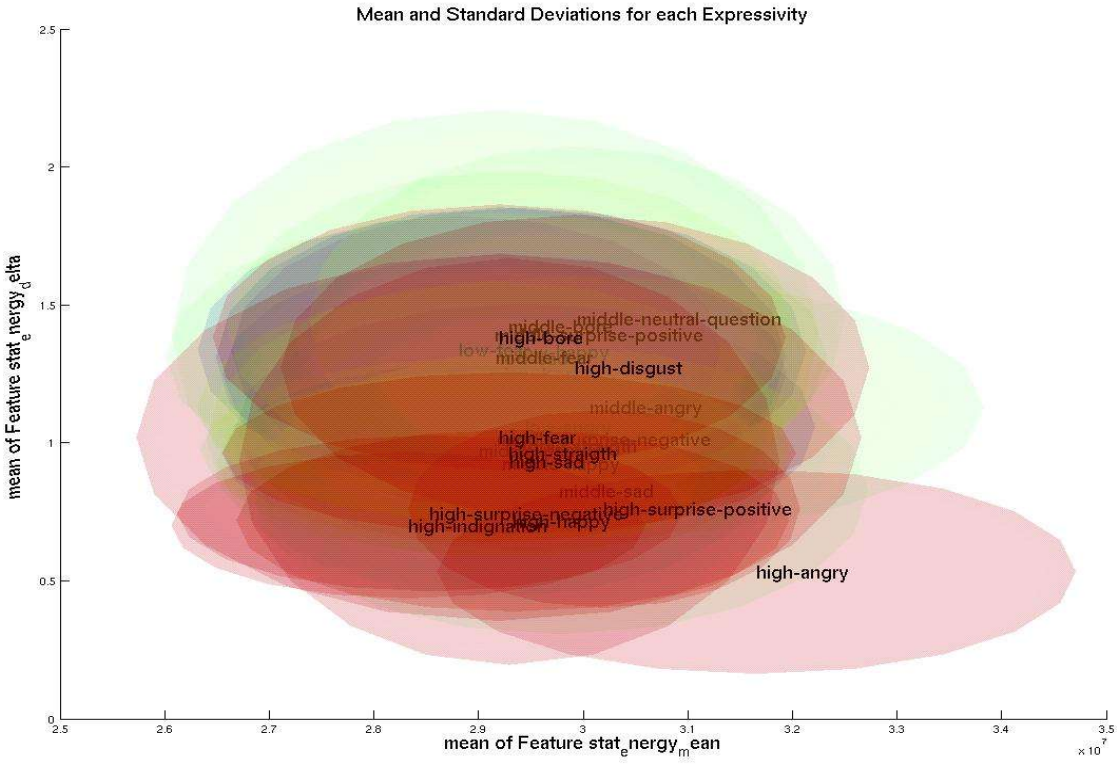
La perturbation de l'énergie est un indice important (voir Partie IV.A). De manière à évaluer son importance, nous traçons (en ordonnée) la moyenne fréquentielle du spectre de l'énergie en fonction de la moyenne fréquentielle du spectre de la fréquence fondamentale (jitter en abscisse). La croissance de la moyenne du spectre de l'énergie signifie la présence de hautes fréquences dans l'intensité dues à des perturbations d'origine physiologique. La forte présence de shimmer et de jitter dans la peur fait de cet espace de descripteurs, un espace discriminant pour cet affect. Les résultats concordent avec la littérature



A l'inverse, l'ennui se distingue par de très faible taux de jitter et de shimmer. Cela traduit l'absence de perturbations d'origine physiologique et une continuité accrue de la fréquence fondamentale et de l'énergie dans l'expression de cet affect. Ces considérations concordent tout à fait avec la littérature.

Une fois l'énergie calculée et moyennée sur l'horizon temporel de chaque phone, nous avons comparée ces moyennes entre elles par classes de phonème. Comme nous le disions dans la partie II-4, l'estimation de l'énergie instantanée, donnée par l'analyse YIN ne se révèle pas d'elle-même, être un bon indice de l'intensité prosodique car elle est très dépendante du contexte phonétique.

Pour obtenir un indice de la variation d'intensité perceptive au cours de la phrase, nous faisons une moyenne statistique des moyennes des énergies par classes de phonèmes (en abscisse de la figure suivante). Par exemple, on prend toutes les énergies moyennes des phones « A » de notre base de données, et on fait la moyenne. Si la moyenne de l'énergie d'un « A » et au-dessus de la moyenne des énergies de sa classe (appelée dorénavant moyenne relative), cela veut dire que ce « A » aura été prononcé avec une emphase d'intensité (appelée alors déviation relative). Cette variation correspond à une stratégie prosodique indépendante du contexte phonétique. Sur la figure suivante, nous avons donc tracer pour chaque catégorie d'affect, la moyenne des moyennes relatives des phones en abscisse et la moyenne des déviations relatives des phones en ordonnée.



On peut conclure que cette mesure est finalement peu informative compte tenue des fortes variances présentées. Seule la moyenne de l'énergie de la colère forte se distingue et nous renseigne sur le fait qu'elle est exprimée plus énergiquement que les autres affects.

Ainsi on peut conclure que l'énergie n'est pas un bon descripteur de l'intensité perçue dans la parole. Dans notre cas, ce n'est pas surprenant car toutes les phrases ont été normalisées en amplitude. Cependant, même si l'on écoute une phrase de colère jouée à faible niveau sonore, on est capable de bien la catégoriser. Ce qui veut dire que l'intensité perçue est relative et non pas absolue. Si la colère se distingue, c'est parce que le signal émis présente une forte quantité de hautes fréquences. Il est donc vraisemblable que la répartition d'énergie dans le spectre est plus informative que l'énergie globale. C'est pourquoi nous allons tenter de mettre en place des descripteurs du timbre (brillance...) pour améliorer les distinctions entre catégories émotionnelles.

V.7.Qualité Vocale:

Pour le moment, aucun descripteur du timbre n'a été importé dans la base de données. Il n'y a pas à l'heure de la rédaction de ce rapport, de résultat à présenter. Cependant, nous travaillons actuellement à la mise en place de descripteurs musicaux du timbre pour caractériser les segments de parole. Ces descripteurs font partie du projet CUIDADO et leur nomenclature est donnée dans (Peeters, 2004) et (Schwarz, 2004) (voir tableau suivant). Grâce à une analyse additive, de nombreux descripteurs harmoniques vont nous permettre de caractériser le timbre (voix soufflée...) en plus des descripteurs spectraux. Nous utiliserons les descripteurs dynamiques (non globaux) présentés dans le tableau suivant:

Liste des descripteurs	Dimension	Acronymes
Energy Features		
Total energy	1	DEi_tot_v
Total harmonic energy	1	DEi_harmo_v
Total noise energy	1	DEi_noise_v
Spectral Features		
Spectral Shape		
Spectral centroid	6	DSi_sc_m
Spectral spread	6	DSi_ss_m
Spectral skewness	6	Dsi_skew_m
Spectral kurtosis	6	Dsi_kurto_v
Spectral slope	6	Dsi_slope_v
Spectral decrease	1	Dsi_decs_c
Spectral rolloff	1	Dsi_rolloff_v
Spectral variation	3	Dsi_variation_v
Harmonic Features		
Fundamental frequency	1	DHi_f 0_v
Fundamental fr. Modulation (frequency, amplitude)	2	f0 Mod AM, FR
Noisiness	1	DHi_noisiness_v
Inharmonicity	1	DHi_inharmo_v
Harmonic Spectral Deviation	3	DHi_devs_v
Odd to Even Harmonic Ratio	3	Dhi_oeratio_v
Harmonic Tristimulus	9	Dhi_tri_v
Harmonic Spectral Shape		
HarmonicSpectral centroid	6	DHi_sc_m
HarmonicSpectral spread	6	DHi_ss_m
HarmonicSpectral skewness	6	DHi_skew_m
HarmonicSpectral kurtosis	6	DHi_kurto_v
HarmonicSpectral slope	6	DHi_slope_v
HarmonicSpectral decrease	1	DHi_decs_c
HarmonicSpectral rolloff	1	DHi_rollof f _v
HarmonicSpectral variation	3	DHi_v ariation_v
Perceptual Features		
Loudness	1	DPi_loud_v
Relative Specific Loudness	24	DPi_specloud_m
Sharpness	1	DPi_sharp_v
Spread	1	DPi_spread_v

Liste des descripteurs	Dimension	Acronymes
Perceptual Spectral Envelope Shape		
Perceptual Spectral centroid	6	DPi_sc_m
Perceptual Spectral spread	6	DPi_ss_m
Perceptual Spectral skewness	6	DPi_skew_m
Perceptual Spectral kurtosis	6	DPi_kurto_v
Perceptual Spectral Slope	6	DPi_slope_v
Perceptual Spectral Decrease	1	DPi_decs_c
Perceptual Spectral Rolloff	1	DPi_rollof f_v
Perceptual Spectral Variation	3	DPi_v ariation_v
Odd to Even Band Ratio	3	DP_ioeratio_v
Band Spectral Deviation	3	DPi_devs_v
Band Tristimulus	9	DPi_tri_v
Various features		
Spectral flatness	4	DPi_sfm_m
Spectral crest	4	DPi_scm_m

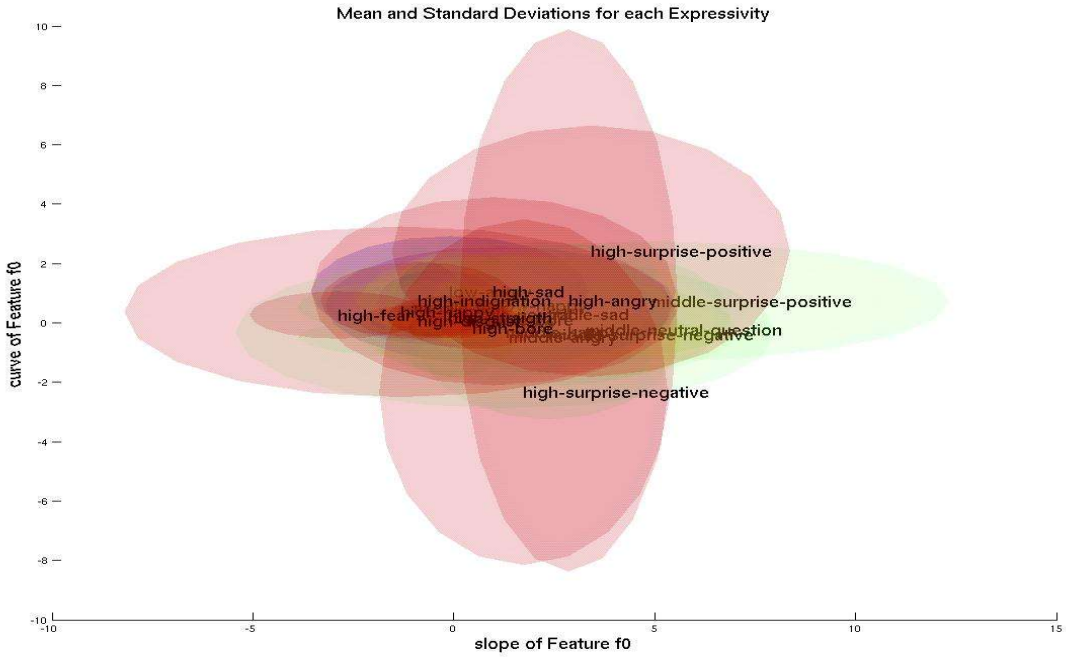
V.8.Tableau récapitulatif des résultats:

Nous présentons ici un tableau récapitulatif de nos diverses remarques sous le même aspect que ceux rencontrés dans la littérature. Pour faciliter la comparaison, nous avons mis en gras des résultats contradictoires avec ceux présentés par la majorité des auteurs. Ces contradictions ne sont pas d'ordre méthodologiques, mais sont plutôt due à la différence du corpus étudié. En effet, il existe très certainement plusieurs stratégies prosodiques pour véhiculer le même affect. Par exemple, notre acteur a exprimé la tristesse en prenant une voix d'enfant pleurnichant; Ceci, dit-il, car l'expression de la tristesse chez l'adulte est très rare pour des raisons évidemment sociales et elle est beaucoup plus évidente chez l'enfant. Cette stratégie prosodique de l'imitation conduit à des résultats très différents de ceux de la littérature (la tristesse possède généralement une fréquence fondamentale très basse).

	f0	Intensité	Débit	Remarques
Colère	- Moyenne ++ - Variabilité ++ - Contours montants - Jitter ++	- Intensité ++ - shimmer +	- Débit ++ - Variabilité -- - Pause -- - Respiration -- - Accélération +	
Peur	- Moyenne + - **Variabilité +** - **Contours descendants** - Jitter ++	- shimmer ++	- Débit ++ - Variabilité -- - Pause ++ - Respiration -- - Accélération 0	- Forte présence de Jitter et Shimmer
Dégoût	- Moyenne -- - **Variabilité --** - Contours descendants - Jitter +	- shimmer -	- Débit - - Variabilité + - Pause 00 - Respiration 00 - Accélération 0	
Joie	- Moyenne ++ - Variabilité ++ - Contours descendants - Jitter 00	- shimmer 00	- **Débit --** - Variabilité ++ - Pause -- - Respiration ++ - Accélération --	- Fortes variations du débit due à l'allongement de voyelles (vers le chant)
Tristesse	- **Moyenne ++** - Variabilité + - **Maximum ++** - **jitter ++**	- **shimmer +**	- Débit -- - Variabilité - - Pause ++ - Respiration - - Accélération ++	
Ennui	- Moyenne -- - Variabilité -- - Jitter --	- shimmer --	- Débit - - Variabilité - - Pause ++ - Respiration 00 - Accélération --	- totale absence de shimmer et jitter
Indignation	- Moyenne + - Variabilité - - Jitter 00	- shimmer +	- Débit 00 - Variabilité 00 - Pause -- - Respiration ++ - Accélération ++	

	f0	Intensité	Débit	Remarques
Surprise Positive	- Moyenne ++ - Variabilité ++ - Jitter 00	- shimmer +	- Débit -- - Variabilité + - Pause ++ - Respiration -- - Accélération 0	- Temps d'explosion des plosives allongées
Surprise négative	- Moyenne ++ - Variabilité ++ - Jitter 00	- shimmer 00	- Débit -- - Variabilité - - Pause - - Respiration ++ - Accélération 0	

On voit en vue des deux dernières lignes de ce tableau que la surprise positive et la surprise négative sont très similaires (pas très surprenant). De manière à les discriminer, nous traçons ici les moyennes des courbures de la fréquence fondamentale de chaque phone. Pour la surprise positive (courbure globalement positive), il semble que la fréquence fondamentale a tendance à évoluer « en creux ». A l'inverse, la surprise négative (courbure globalement négative) présente des phones dont la fréquence fondamentale a tendance à évoluer « en bosse ».



Conclusion:

Cette première étude nous a permis d'aborder certains points primordiaux de l'analyse des émotions. Tout d'abord, elle a été l'occasion de se confronter à la littérature du domaine qui possède la particularité d'être soit « antique »: Des théorie des émotions sont nées dans de très nombreuses civilisations et à travers tous les âges); Soit « moderne »: Regard nouveau depuis l'apparition des sciences expérimentales dans les sciences humaines (psycho acoustique, psycholinguistique...)).

Puis, elle nous a donné l'occasion de constituer une base de données. Sa mise en place s'est révélée assez longue et nous a permis d'entrevoir les difficultés que pose une telle entreprise. Ainsi, nous sommes aujourd'hui en possession d'une base saine, segmentée et analysée. Elle est assez grande comparée à quelques unes rencontrées dans la littérature (issues de la segmentation manuelle). La prochaine étape dans la constitution de cette base est sa validation par un corpus d'auditeurs lors de tests perceptifs (voir Travaux Futurs).

Nous nous sommes focalisés sur l'importance de la prosodie dans la communication des affects (partie cognitive) pour la simple raison que notre acteur a simulé ces affects. Pour cela, nous avons notamment créé un descripteur de débit de parole dynamique non recensé pour le moment dans la littérature. Nous avons mis en évidence qu'un tel descripteur possède une importance capitale dans la caractérisation de la joie et donc dans sa distinction avec la colère (problème généralement cité dans la littérature). Ce descripteur dynamique participe à l'outil de synthèse de parole à partir du texte.

En ce qui concerne l'évolution de la fréquence fondamentale, nos résultats corroborent tout à fait à l'opinion générale, ce qui valide en quelque sorte la méthode employée. Les différences de résultats (peur et tristesse) nous amènent à penser que l'acteur a utilisé des stratégies prosodiques particulières et qu'il serait intéressant de considérer d'autres stratégies pour leurs diversités.

Dans l'étude de l'énergie, nous avons vite cerné une limitation majeure qui oriente aujourd'hui notre questionnement sur la répartition spectrale de cette énergie, caractéristique de la qualité vocale. Comme en musique, la paramétrisation du timbre semble être primordiale. Nous allons donc tenter de mettre l'accent sur la définition de tels paramètres. Nous considérerons les déviations par rapport aux valeurs calculées sur les phrases neutres comme indices prosodiques. D'ores et déjà, nous savons l'importance de tes descripteurs par l'observation des spectrogrammes, par l'audition des phrases synthétisées ou transformées et par la lecture de la littérature à ce sujet.

Pour finir, il faut souligner que la majeure partie du travail réalisé durant ce stage réside dans la construction, la maintenance et l'élaboration de l'outil d'exploration de la parole, de transformation et de synthèse nommé Talkapillar. Deux présentations de cet outil réunissant plusieurs « briques » (Matlab, PostgreSQL, Perl, SVP, Additive...) ont été faites durant ce stage: Pour un séminaire Recherche et Création (IRCAM) et pour le groupe de travail sur la Voix (IRCAM). Une troisième est envisagée pour la soutenance malgré sa courte durée. Cet outil est devenu peu à peu la base de données idéale pour réaliser des opérations statistiques et des analyses rapides de corpus et sera donc utiliser pour mener à bien nos travaux futurs.

Travaux futurs:

La conclusion du rapport appuie sur l'importance de la description du timbre. Nous allons ainsi tenter de créer de tels descripteurs. Une technologie employable est celle de la conversion de voix appelée « voice morphing ». Par apprentissage sur deux voix distinctes prononçant la même chose, on est capable d'apprendre la transformation de l'une à l'autre et ainsi d'en synthétiser une, uniquement à partir de l'autre. Nous assumons ici une hypothèse forte que nous tenterons de développer durant la thèse: Le problème du changement d'identité vocale peut être assimilé au problème de changement d'état émotionnel intra locuteur. Cela veut dire que la synthèse d'affects et le changement d'identité nécessitent les mêmes outils.

Si ces deux opérations sont réunissables dans le même champ de travail, cela peut lui conférer une grande flexibilité et une élargir sont champ de résultats. En effet, si la conversion d'une voix A à une voix B se fait par comparaisons de phrases neutres de deux locuteurs A et B, et que le changement d'état émotionnel chez un locuteur A se fait par comparaison entre des phrases neutres A-N à des phrases affectées quelconque A-Q, alors on est en mesure de penser que par combinaison des deux transformations, on puisse synthétiser pour le locuteur B, tous les affects du locuteurs A, sans qu'il n'est jamais prononcé de phrases affectées. Cette hypothèse a déjà été appuyée dans la rédaction d'un article soumis à la conférence IEEE « special issue on Expressive Speech Synthesis », intitulé « Multi-Speaker ESS with Only One Expressive Database » (Appendix C).

Cette hypothèse pose une question sous-jacente importante: Peut-on quantifier l'information vocale objectivement en prenant en compte l'aspect contextuel et cognitif de chaque individu? Si nous avons toujours exprimé les émotions primaires de manière « irrégulière » ou non neutre, c'est pour donner de l'information supplémentaire au message. Conséquemment, Nous tenterons d'introduire la théorie de l'information dans les analyses psycho acoustiques par le biais de filtres cognitifs.

Devant la diversité des stratégies prosodiques selon les émotions, nous préconisons la mise en place de sous-ensembles de descripteurs pertinents. Choisis pour l'instant de façon heuristique, le nombre croissant de descripteurs (descripteurs du timbre nombreux) nous oblige à employer d'autres méthodes. C'est pourquoi nous allons, dans nos futurs travaux, mettre en place des algorithmes de classification automatique par apprentissage. La partie recherche de la mise en place de tels algorithmes résidera dans la « non linéarité » de l'espace de descripteurs en fonction des affects (sous-ensembles discriminants différents maximisant les distances entre affects).

Des tests d'écoutes vont être mis en place de manière à valider la base de données et à évaluer nos synthèses. En mélangeant des synthèses à des phrases du corpus, nous allons étudier leurs catégorisations sous trois angles. Le premier tente de déterminer l'ensemble des phrases du corpus les mieux catégorisées. La suite de nos investigations s'appuiera ainsi sur ce corpus filtré dont le contenu sera étiqueté par les sujets du test. Le second angle est l'évaluation de la même manière de la catégorisation des phrases synthétisées pour voir dans quelle mesure celles-ci répondent à nos attentes. Enfin, le troisième angle vise à déterminer l'effet de la synthèse sur la catégorisation par une mesure du temps de réponse des sujets. Cela pour vérifier si la synthèse est détectée par les sujets et si elle leur pose un problème pour la reconnaissance des affects.

D'autres investigations sont possibles et notamment dans le domaine de la musique (Beller, 2005). Quelques unes ont d'ores et déjà vu le jour et ont été le sujet d'un article (Beller et al., 2005) et d'une conférence menée au JIM, début juin 2005, intitulée « A hybrid concatenative synthesis system on the intersection of music and speech » (Appendix B). Des expériences vont être menée sur le Sprechgesang dans le cadre du futur opéra de Emmanuel Nunez (2007). Une implémentation temps réel du système de synthèse et de segmentation automatique (alignement) est envisagée par l'équipe ATR de l'IRCAM.

Références:

Abadjieva *et al.*, 1993

E. ABADJIEVA, I. R. MURRAY, J. L. ARNOTT.
Applying analysis of human emotional speech to enhance synthetic speech.
Eurospeech, 909-912, Berlin, Allemagne, 1993.

Auchlin Simon, 2004

A. AUCHLIN A.-C. SIMON.
Gabarits prosodiques, empathie(s) et attitudes.
Cahiers de l'Institut de Linguistique de Louvain, 30, 181-206, 2004.

Beller, 2004

G. BELLER.
Un synthétiseur vocal par sélection d'unités.
Rapport de stage, IRCAM, 2004.

Beller, 2005

G. BELLER.
La musicalité de la voix parlée.
Maitrise de musique, Université Paris 8, Paris, 2005.

Beller et al., 2005

Gregory BELLER, Diemo SCHWARZ, Thomas HUEBER, and Xavier RODET.
A hybrid concatenative synthesis system on the intersection of music and speech.
JIM, IRCAM, 2005. Analysis/synthesis team.

Bezooijen, 1984

R. Van BEZOOIJEN.
Characteristics and recognizability of vocal expressions of emotion.
Dordrecht, Pays-Bas, 1984.

Black, 2003

A.W. BLACK.
Unit selection and emotional speech.
Eurospeech, 2003.

Bonner, 1943

M. R. BONNER.
Changes in the speech pattern under emotional tension.
American Journal of Psychology, 56:262-273, 1943.

Bosh, 2000

L.T. BOSH.
Emotions: what is possible in the asr framework?
ISCA Workshop on Speech and Emotion, 2000.

Bulut *et al.*, 2002

M. BULUT, S. SHRIKANTH, S.S. NARAYANAN, A. K. SYRDAL.
Expressive speech synthesis using a concatenative synthesizer.
ICSLP, ATT Labs-Research, Florham Park, NJ, 2002.

Burkhardt Sendlmeier, 2000
F. BURKHARDT W.F. SENDLMEIER.
Verification of acoustical correlates of emotional speech using formant-synthesis.
ISCA Workshop on Speech and Emotion, Northern Ireland, 151-156, 2000.

Bänziger, 2004
T. BÄNZIGER.
Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles.
Psychologie, Faculté de psychologie et des sciences de l'éducation de l'Université de Genève, Genève, 2004.
Sous la direction de Klaus R. Scherer.

Cahn, 1989
J.E. CAHN.
Generating expression in synthesized speech.
Master's thesis, MIT, 1989.

E. Cahn, 1990
Cahn J. E..
The generation of affect in synthesized speech.
Journal of the American Voice I/O Society, 1-19, 8, July 1990.

Chung, 2000
S.-J. CHUNG.
L'expression et la perception de l'émotion extraite de la parole spontanée: évidences du coréen et de l'anglais.
Phonétique, Université PARIS III - Sorbonne Nouvelle: Institut de Linguistique et Phonétique Générales et Appliquées, Paris, 2000.
sous la direction de J. Vaissière.

Cutler *et al.*, 1986
A. CUTLER, J. MEHLER, D. NORRIS, J. SEGUI.
The syllable's differing role in the segmentation of french and english, Journal of Memory and Language, 385-400.
http://www.nici.kun.nl/Publications/2002/16340.html, 1986.

De Mareüil *et al.*, 2000
P. Boula de MAREÜIL, P. CÉLÉRIER, J. TOEN.
Generations of emotions by a morphing technique in english, french and spanish.
2000.
Elan TTS and LIMSI-CNRS.

Devillers *et al.*, 2003a
L. DEVILLERS, L. LAMEL, I. VASILESCU.
Emotion detection in task-oriented spoken dialogs.
IEEE ICME, 2003.

Devillers *et al.*,
L. DEVILLERS, I. VASILESCU, L. LAMEL.
Détection d'émotions dans des dialogues oraux.
www.limsi.fr/RS2003FF/CHM2003/TLP2003/TLP6/tlp6.html.

Devillers *et al.*, 2003b
L. DEVILLERS, I. VASILESCU, C. MATHON.
Prosodic cues for perceptual emotion detection in task-oriented human-human corpus.
ICPhS, 2003.

Drioli *et al.*, 2003
C. DRIOLI, G. TISATO, P. COSI, F. TESSER.
Emotions and voice quality: Experiments with sinsoidal modeling.
VOQUAL'03, 127-132, Laboratory of Phonetics and Dialectology, ISTC-CNR, Institute of Cognitive Sciences and Technology, Italy, August 2003.
drioli@csrf.pd.cnr.it.

Fagyal, 1995
Z. FAGYAL.
Aspects phonostylistiques de la parole médiatisée lue et spontanée: Age, prestige, situation, style et rythme de parole de l'écrivain M. Duras.
PhD thesis, Université de la Sorbonne Nouvelle, Paris, 1995.

Fairbanks Pronovost, 1938
G. FAIRBANKS W. PRONOVOST.
Vocal pitch during simulated emotion.
Science, 88:382-383, 1938.

Fónagy, 1983
I. FÓNAGY.
La vive voix: essais de psycho-phonétique.
1983.

Fónagy Bérard, 1972
I. FÓNAGY E. BÉRARD.
Il est huit heure: Contribution à l'analyse sémantique de la vive voix.
Phonetica, 26:157-192, 1972.

Galarneau *et al.*,
A. GALARNEAU, P. TREMBLAY, P. MARTIN.
Dictionnaire de la parole.
http://www.lli.ulaval.ca/labo2256/lexique/dico.html.
Laboratoire de Phonétique et Phonologie de l'Université Laval à Québec.

Henton, 1997
C.G. HENTON.
Method and apparatus for automatic generation of vocal emotion in a synthetic text-to-speech system.
United States Patent 5,860,064, Apple Computer, Inc. (Cupertino, CA), February 24 1997.

Horri, 1982
Y. HORRI.
Jitter and shimmer differences among sustained vowel phonations.
Journal of Speech and Hearing Research, 25:12-14, 1982.

Hutter, 1968
G. L. HUTTER.

Relations between prosodic variables and emotions in normal american english utterances.
Journal of Speech and Hearing Research, 11:481- 487, 1968.

Jiang *et al.*, 2000
D.-N. JIANG, W. ZHANG, L.-Q. SHEN, L.-H. CAI.
prosody analysis and modeling for emotional speech synthesis.
2000.

Juslin Laukka, 2003
P.N. JUSLIN P. LAUKKA.
Communication of emotions in vocal expression and music performance: Different channels, same code?
Psychological Bulletin, 129(5), 770-814, 2003.

Kaiser, 1962
L. KAISER.
Communication of affect by single vowels.
Synthese, 14:300-319, 1962.

Klabbers Santen, 2002
E. KLABBERS J.P.H. Van SANTEN.
Clustering of foot-based pitch contours in expressive speech.
ICSLP Speech Synthesis Workshop, Pittsburg, Center for Spoken Languiage Understanding, 2002.

Klingholz Martin, 1985
F. KLINGHOLZ F. MARTIN.
Quantitative spectral evaluation of shimer and jitter.
Journal of Speech and Hearing Research, 28:169-174, 1985.

Labov, 1972
W. LABOV.
Langugage in the inner city : Studies in the black english vernacular.
University of Pennsylvania Press, Philadelpia, 1972.

Lacheret-Dujour Beaugendre, 1999
A. LACHERET-DUJOUR F. BEAUGENDRE.
1999.

Laukka, 2004
P. LAUKKA.
Vocal Expression of Emotion: discrete-emotions and dimensional Accounts.
PhD thesis, Uppsala, 2004.

Laukkanen *et al.*, 1996
A.M. LAUKKANEN, E. VILKMAN, P. ALKU, H. OKSANEN.
Physical variations related to stress and emotional state: A preliminary study.
Journal of Phonetics, 24:313-335, 1996.

Lee *et al.*, 2002
C. LEE, S.S. NARAYANAN, R. PIERACCINI.
Recognition of negative emotions from the speech signal.

Automatic Speech Recognition and Understanding, 2002.

Leinonen *et al.*, 1997
L. LEINONEN, T. HILTUNEN, I. LINNANKOSKI, M.-L. LAAKSO.
Expression of emotional-motivational connotations with a one-word utterance.
J. Acoust. Soc. Am., 102(3):1853-1863, 1997.

Lida *et al.*, 2003
A. LIDA, N. CAMPBELL, F. HIGUCHI, M. YASUMURA.
A corpus-based speech synthesis system with emotion.
Speech Commun., 40(1-2):161-187, 2003.

Lieberman, 1961
P. LIEBERMAN.
Perturbations in vocal pitch.
J. Acoust. Soc. Am., 33:597-603, 1961.

Lieberman Michaels, 1962
P. LIEBERMAN S. B. MICHAELS.
Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech.
Journal of the Acoustical Society of America, 34(7), 922-927, 1962.

Litman. Forbes, 2003
D. LITMAN. K. FORBES.
Recognizing emotions from student speech in tutoring dialogues.
IEEE Automatic Speech Recognition and Understanding Workshop ASRU, 2003.

Léon, 1970
P.R. LÉON.
Systématique des fonctions expressives de l'intonation.
Studia Phonetica, 3:57-74, 1970.

Miller, 1987
J.L. MILLER.
Mandatory processing in speech perception: A case study, Modularity in knowledge representation and natural-language understanding, 310-322.
J.L. Garfield, www.pallier.org/papers/thesis/thesehtml/node8.html, 1987.

Morel Bänziger, 2004
M. MOREL T. BÄNZIGER.
Le rôle de l'intonation dans la communication vocale des émotions : test par la synthèse.
Cahiers de l'Institut de Linguistique de Louvain (CILL), 30, 207-232, 2004.

Mozziconacci, 1998
S.J.L. MOZZICONACCI.
Speech Variability and Emotion: Production and Perception.
PhD thesis, Technical University Eindhoven, 1998.

Murray. Arnott, 1995
I.R. MURRAY. J.L. ARNOTT.
Implementation and testing of a system for producing emotion-by-rule in synthetic speech.

Speech Commun., 16(4):369-390, 1995.

Murray Arnott, 1993
I. R. MURRAY J. L. ARNOTT.
Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion.
Journal of the Acoustical Society of America, 93:1097-1108, 1993.

Murray *et al.*, 2000
I. R. MURRAY, M. D. EDGINGTON, D. CAMPION, J. LYNN.
Rule-based emotion synthesis using concatenated speech.
ISCA Workshop on Speech and Emotion, A conceptual framework for research, Northern Ireland, Bell T. lab, 2000.

Narayanan *et al.*, 2002
S. NARAYANAN, C. LEE, R. PIERACCINI.
Combining acoustic and language information for emotion recognition.
ICSLP, 2002.

Oudeyer, 2002
P.-Y. OUDEYER.
The production and recognition of emotions in speech: features and algorithms.
International Journal of Human Computer Interaction, 2002.

Oudeyer, 2003
P.-Y. OUDEYER.
The production and recognition of emotions in speech: features and algorithms.
Human-Computer Studies, 59:157 183, Nov 2003.
Sony CSL Paris, 6, rue Amyot, 75005 Paris, France.

Peeters, 2004
G. PEETERS.
A large set of audio features for sound description (similarity and classication) in the cuidado project.
Technical Report, IRCAM, 2004

Pereira, 2000
C. PEREIRA.
Perception and expression of emotions in speech.
PhD thesis, Macquarie University, 2000.

Pereira Watson, 1998
C. PEREIRA C. WATSON.
Some acoustic characteristics of emotion.
Fifth International Conference on Spoken Language Processing, Sydney, 1998.

Petrushin, 1999
V.A. PETRUSHIN.
System, method and article of manufacture for an emotion detection system improving emotion recognition.
United States Patent 6,353,810, Accenture LLP (Palo Alto, CA), August 31 1999.

Picard, 1997
R. PICARD.
Affective Computing.
1997.

Rouas *et al.*, 1998
J.-L. ROUAS, J. FARINAS, F. PELLEGRINO.
Evaluation automatique du débit de parole sur des données multilingues spontannées.
1998.

Samal Iyengar, 1992
A. SAMAL P. IYENGAR.
Automatic recognition and analysis of human faces and facial expression: a survey.
Pattern Recognition, 25 (1), 65-77, 1992.

Scherer, 1989
K.R. SCHERER.
Vocal correlates of emotion, 165-197.
1989.

Scherer, 2003
K.R. SCHERER.
Vocal communication of emotion: a review of research paradigms.
Speech Commun., 40(1-2):227-256, 2003.

Scherer Oshinsky, 1977
K.R. SCHERER OSHINSKY.
Cue utilization in emotion attribution from auditory stimuli.
Motivation and Emotion, 1(4):331-346, 1977.

Schlossberg, 1954
H. SCHLOSSBERG.
Three dimensions of emotion.
Psychological Review, 61, 81-88, 1954.

Schoentgen de Guchtenneere, 1995
J. SCHOENTGEN R. de GUCHTENNEERE.
Time series analysis of jitter.
Journal of Phonetics, 23:189-201, 1995.

Schröder, 2001
M. SCHRÖDER.
Emotional speech synthesis-a review.
Eurospeech, Aalborg, 561-564, DFKI, Saarbrücken, Germany: Institute of Phonetics, University of the Sarland, 2001.
schroed@dfki.de.

Schuller *et al.*, 2005
B. SCHULLER, R.J. VILLAR, G. RIGOLL, M. LANG.
Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition.
ICASP, schuller@ei.tum.de, 2005. Institute for Human-Machine Communication.

Schwarz, 2000
Diemo SCHWARZ.
A System for Data-Driven Concatenative Sound Synthesis.
pages 97102, Verona, Italy, December 2000.

Schwarz, 2003a
Diemo SCHWARZ.
New Developments in Data-Driven Concatenative Sound Synthesis.
pages 443-446, Singapore, October 2003.

Schwarz, 2003b
Diemo SCHWARZ.
The C AT E R P I L L A R System for Data-Driven Concatenative Sound Synthesis.
pages135140, London, UK, September 2003.

Schwarz, 2004
Diemo SCHWARZ.
Data-Driven Concatenative Sound Synthesis.
These de doctorat, Universite Paris 6 Pierre et Marie Curie, Paris, 2004.
Caterpillar. Web page, 2005. http://recherche.ircam.fr/anasyn/schwarz/thesis.

Shafran Mohri, 2005
I. SHAFRAN M. MOHRI.
a comparison of classifiers for detecting emotion from speech.
IEEE, Center for langage and Speech Processing, zakshafran@jhu.edu, mohri@cs.nyu.edu, 2005.

Skinner, 1935
E.R. SKINNER.
A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness.
Speech Monograph, 2:81-137, 1935.

Streeter *et al.*, 1983
L.A. STREETER, N.H. MACDONALD, W. APPLE, R.M. KRAUSS, K.M. GALOTTI.
Acoustic and perceptual indicators of emotional stress.
J. Acoust. Soc. Am., 73(4):1354-1360, 1983.

Takahashi Koike, 1975
H. TAKAHASHI Y. KOIKE.
Some perceptual dimensions and acoustical correlates of pathologic voices.
Acta Otolaryngologica, Suppl.:338, 1975.

Tomkins, 1984
S.S. TOMKINS.
Approches to Emotion, Affect theory, 163-195.
Lawrence Erlbaum Associates Publishers, Londres, 1984.

Tsuzuki *et al.*, 2004
R. TSUZUKI, H. ZEN, K. TOKUDA, T. KITAMURA, M. BULUT, S.S. NARAYANAN.
Constructing emotional speech synthesizers whith limited speech database.
ICSLP, Department of Computer Science and Engineering, Nagoya Istitute of Technology, 2004.

Vasilescu Devillers, 2003
I. VASILESCU L. DEVILLERS.
Lrec 2004, speech prosody 2004.
ICPhS, 2003.

Vasilescu *et al.*, 2004
I. VASILESCU, L. DEVILLERS, C. CLAVEL, EHRETTE.
Base de données de fiction pour la détection d'émotion dans des situations anormales.
Interspeech-2004, 2277-2280, 2004.

Wheeldon, 1994
L. WHEELDON.
Do speakers have access to a mental syllabary, Cognition, 50, 239-259.
www.pallier.org/papers/thesis/thesehtml/node8.html, 1994.

Williams Stevens, 1972
U. WILLIAMS K.N. STEVENS.
Emotions and speech: some acoustical correlates.
JASA, 52, 1238-1250, 1972.

Yac, 2003
Recognition of Emotions in Interactive Voice Response Systems, 8th European Conference on Speech Communication and Technology, Hewlett-Packard laboratories Palo Alto, July 2003.
yacoub@hp.com.

Zellner, 1998
B. ZELLNER.
Caractérisation du débit de parole en français.
Journées d'Etude sur la Parole, Brigitte.zellner@imm.unil.ch, 1998. Faculté de Lettres, Université de Lausanne.

Appendix A:

De manière à préciser des descripteurs rencontrés jusqu'ici, nous établissons une liste détaillée dont la nomenclature est extraite de (Pereira, 2000). Celle-ci souligne l'importance de la taille de la fenêtre d'analyse qu'elle prend 4,5 fois plus longue que la période fondamentale.

Nom	Description	Calcul
Énergie		
ENER_MAX	Maximum de l'énergie	
ENER_MAX_POS	Position du Maximum d'énergie	
ENER_MIN	Minimum de l'énergie	
ENER_MIN_POS	Position du Minimum d'énergie	
ENER_REG_COEF	Coefficient de la droite issue de la régression linéaire sur la courbe d'énergie pour toute la phrase	$EnergyRegCoef = \frac{S_{ene,xy}}{S_{ene,x}}$ $S_{ene,xy} = \sum_{i=1}^{N} i \cdot E_i - \frac{1}{N} \cdot \sum_{i=1}^{N} i \cdot \sum_{i=1}^{N} E_i$ $S_{ene,x} = \sum_{i=1}^{N} i^2 - \frac{1}{N} \cdot \left(\sum_{i=1}^{N} i \right)^2$
ENER SQR_ERR	Erreur quadratique moyenne sur l'estimation du coefficient de régression	$EneSqrErr = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(E_i - \left(\mu_E - \frac{S_{ene,xy}}{S_{ene,x}} \cdot \mu_i \right) - \frac{S_{ene,xy}}{S_{ene,x}} \cdot i \right)^2$ $\mu_E = \frac{1}{N} \cdot \sum_{i=1}^{N} E_i$ $\mu_i = \frac{1}{N} \cdot \sum_{i=1}^{N} i$
ENER_MEAN:	Moyenne de l'énergie	$MeanEne = \frac{\sum_{i=1}^{N} E_i}{N}$

ENER_ VAR	Variance de l'énergie	$VarEne = \frac{\sum_{i=1}^{N} E_i^2}{N} - \mu_E^2$ μ_E ² = Energy mean
Fréquence Fondamentale		
f0_MAX	f0 maximum	
f0_MAX_POS :	Position du maximum de f0	
f0_MIN	f0 minimum	
f0_MIN_POS	Position du minimum de f0	
f0_REG_COEF	Coefficient de la droite issue de la régression linéaire sur la courbe de f0 pour toute la phrase	$F0RegCoef = \frac{S_{F0,xy}}{S_{F0,x}}$ $S_{F0,xy} = \sum_{i=1}^{N} i \cdot F0_i - \frac{1}{N} \cdot \sum_{i=1}^{N} i \cdot \sum_{i=1}^{N} F0_i$ $S_{F0,y} = \sum_{i=1}^{N} i^2 - \frac{1}{N} \cdot \left(\sum_{i=1}^{N} i \right)^2$
f0_SQR_ERR	Erreur quadratique moyenne sur l'estimation du coefficient de régression	$F0SqrErr = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(F0_i - \left(\mu_{F0} - \frac{S_{F0,xy}}{S_{F0,x}} \cdot \mu_i \right) - \frac{S_{F0,xy}}{S_{F0,x}} \cdot i \right)^2$ $\mu_{F0} = \frac{1}{N} \cdot \sum_{i=1}^{N} F0_i$ $\mu_i = \frac{1}{N} \cdot \sum_{i=1}^{N} i$
f0_MEAN:	Moyenne de f0	$MeanF0 = \frac{\sum_{i=1}^{N} F0_i}{N}$
f0_VAR	Variance de f0	$VarF0 = \frac{\sum_{i=1}^{N} F0_i^2}{N} - \mu_{F0}^2$ μ_{f0} ² = Pitch mean

Jitter.	Perturbation de f0	$jitter = \frac{\sum_{i=2}^{N-1} \left\| 2 \cdot T_i - T_{i-1} - T_{i+1} \right\|}{\sum_{i=2}^{N-1} T_i}$
Voisé/non-voisé		
f0_FIRST_VCD_FRAME	Valeur de f0 pour la première trame voisée	
f0_LAST_VCD_FRAME	Valeur de f0 pour la dernière trame voisée	
NUM_VOICED_REGIONS.	Nombre de régions contenant plus de trois trames voisées successives	
NUM_UNVCD_REGIONS	Nombre de régions contenant plus de trois trames non-voisées successives	
NUM_VOICED_FRAMES	Nombre de trames voisées	
NUM_UNVCD_FRAMES	Nombre de trames non- voisées	
LGTH_LNGST_V_REG	Longueur de la plus longue région voisée	
LGTH_LNGST_UV_REG	Longueur de la plus longue région non-voisée	
RATIO_V_UN_FRMS	Rapport entre le nombre de trames voisées et le nombre de trames non-voisées	$RatVcdUnvc\ dFrms = \frac{\sum voiced\_frames}{\sum unvoiced\_frames}$
RATIO_V_UN_REG	Rapport entre le nombre de régions voisées et le nombre de régions non-voisées	$RatVcdUnvc\ dRg = \frac{\sum voiced\_regions}{\sum unvoiced\_regions}$
RATIO_V_ALL_FRMS	Rapport entre le nombre de trames voisées et le nombre total de trames	$RatVcdAllF\ rms = \frac{\sum voiced\_frames}{N}$
RATIO_UV_ALL_FRMS	Rapport entre le nombre de trames non-voisées et le nombre total de trames	$RatUnvcdAl\ lRg = \frac{\sum unvoiced\_regions}{N}$
Dérivée de f0		
f0_DER_MAX.	Max de la dérivée de f0	
f0_DER_MAX_POS	Position du maximum de la dérivée de f0	

f0_DER_MIN	Min de la dérivée de f0	
f0_DER_MIN_POS	Position du minimum de la dérivée de f0	
f0_DER_REG_COEF	Coefficient de la droite issue de la régression linéaire sur la courbe de la dérivée de f0 pour toute la phrase	
f0_DER_SQR_ERR	Erreur quadratique moyenne sur l'estimation du coefficient de régression	
f0_DER_MEAN	Moyenne de la dérivée de f0	
f0_DER_VAR	Variance de la dérivée de f0	
Descripteurs Haut-Niveau: **N = nombre de régions voisées**		
f0_REG_MEAN	Moyenne des moyennes de f0 de toutes les régions voisées	$F0AbsMean = \frac{\sum_N \overline{F0}_n}{N}$
f0_REG_VAR	Variance des moyennes de f0 de toutes les régions voisées	$VarF0Mean = \frac{\sum_N \left(\overline{F0}_n - F0AbsMean\right)^2}{N}$
f0_REG_MAX_MEAN	Moyenne des maximums de f0 de toutes les régions voisées	$MaxF0Mean = \frac{\sum_N F0_{\max_n}}{N}$
f0_REG_MAX_VAR	Variance des maximums de f0 de toutes les régions voisées	$VarF0Max = \frac{\sum_N \left(F0_{\max_n} - MaxF0Mean\right)^2}{N}$
PEAKS_INCREASING	Pour chaque région voisée, on définit 4 points: - Début - Fin - Premier Max - Second Max On somme toutes les différences entre deux points croissants successifs divisés par leur différence temporelle respective. Puis on fait la moyenne arithmétique sur toutes les régions voisées	$PeaksIncre ase = \frac{\sum_N \left(\sum_{increase} \frac{\left\|F0_i - F0_j\right\|}{t_i - t_j} \right)}{N}$ *i , j = représentent* l'un des quatre points considérés i < j $t_i < t_j$ $f0_i < f0_j$

PEAKS_DECREASING	Pour chaque région voisée, on définit 4 points: - Début - Fin - Premier Max - Second Max On somme toutes les différences entre deux points décroissants successifs divisés par leur différence temporelle respective. Puis on fait la moyenne arithmétique sur toutes les régions voisées	$PeaksDecrease = \frac{\sum_{N}\left(\sum_{decrease}\frac{\lvert F0_i - F0_j\rvert}{t_i - t_j}\right)}{N}$ *i , j = représentent* l'un des quatre points considérés i < j $t_i < t_j$ $f0_i > f0_j$
RANGE_MEAN	Moyenne des intervalles pour les régions voisées	$MeanRange = \frac{\sum_{N}\left(F0_{\max_n} - F0_{\min_n}\right)}{N}$
FLATNESS	Moyenne de toutes les MEAN/MAX des régions voisées (*100)	$Flatness = \frac{\sum_{N}\frac{\overline{F0}_n}{F0_{\max_n}}\cdot 100}{N}$
DUR_PITCH_MAX	Moyenne des durées relatives entre le début d'une région voisée et la position du maximum de f0 (*100)	$DurMaxPitch = \frac{\sum_{N}\left(t_{\max n} - t_{start_n}\right)\cdot 100}{N}$
PEAKS_INCREASE_UTT	Chaque région voisée est définie par son maximum. On somme toutes les différences entre deux points croissants successifs divisés par leur différence temporelle respective. Puis on fait la moyenne arithmétique sur toute la phrase.	$PeaksIncreaseUtt = \sum_{increase}\frac{\lvert F0_{\max_i} - F0_{\max_j}\rvert}{t_i - t_j}$ t_i , t_j = Position des maxima des régions voisées. i < j
PEAKS_DECREASE_UTT	Chaque région voisée est définie par son maximum. On somme toutes les différences entre deux points décroissants successifs divisés par leur différence temporelle respective. Puis on fait la moyenne arithmétique sur toute la phrase.	$PeaksDecreaseUtt = \sum_{decrease}\frac{F0_{\max_i} - F0_{\max_j}}{t_i - t_j}$ t_i , t_j = Position des maxima des régions voisées. i < j
DUR_MEAN_UTT	Moyenne des durées des régions voisées	$DurMean = \frac{\sum_{N} length_n}{N}$
ENER_MEAN_UTT	Moyenne des moyennes de l'énergie divisée par le maximum d'énergie de la phrase	$EnerMean = \frac{\frac{\sum_{N}\overline{E}_n}{N}\cdot 100}{E_{MAX\_UTT}}$

ENER_DUR_START	Moyenne des durées relatives entre le début d'une région voisée et la position du maximum de l'énergie (*100) et divisée par le max d'énergie de la phrase	$EneDurStar\ t = \frac{\sum_N \frac{E_{\max_n}}{t_{\max_n} - t_{start_n}}}{E_{MAX\_UTT}}$
ENER_DUR_END	Moyenne des durées relatives entre la position du maximum de l'énergie et la fin (*100) et divisée par le max d'énergie de la phrase	$EneDurEnd\ = \frac{\sum_N \frac{E_{\max_n}}{t_{end_n} - t_{\max_n}}}{E_{MAX\_UTT}}$
ENER_VEHEMENCE	Moyenne de la véhémence de l'énergie (MEAN/MIN) sur les régions voisées	$EneVehemen\ ce = \frac{\sum_N \frac{\overline{E}_n}{E_{\min}}}{N}$
ENER_FLATNESS	Moyenne des flatness (MEAN/MAX) de l'énergie sur les régions voisées	$EneFlatness = \frac{\sum_N \frac{\overline{E}_n}{E_{\min}}}{N}$
ENER_MAX_RATIO	Relation entre le maximum d'énergie sur toute la phrase et sa position	$EneMaxRati\ o = \frac{E_{MAX\_UTT}}{t_{MAX\_UTT}}$
ENER_MAX_RATION_REG	Relation ente le max de l'énergie d'une région voisée et le max sur toute la phrase divisée par la position du maximum de la région. Puis moyenne arithmétique sur toutes les régions voisées	$EnerMaxRat\ io_{region} = \frac{\sum_N \frac{E_{\max_n} / E_{MAX\_UTT}}{t_{MAX\_UTT}} \cdot 100}{N}$
TREMOR_MEAN	Le tremor est le nombre de passage par zéro de la dérivée de la courbe d'énergie fenêtrée	$TremorMean\ = \frac{\sum_N ene\_tremor_n}{N}$
Qualité Vocale		
MEAN1	Moyenne arithmétique de la valeur (f) d'un paramètre sur toutes les trames (nframes) composant une région voisée	$Mean1(f) = \frac{\sum_{nframes} f_i}{nframes}$
MEAN2	Mean2 est la valeur existante du paramètre la plus proche de Mean1.	$Mean2(f) = \left(\min_i \left\| f_i - \overline{f}_n \right\|\right)$
H2N_MAX	Maximum des Mean2 du rapport Harmonique/Bruit	
H2N_REG	Maximum du rapport Harmonique/Bruit pour chaque région voisée	
H2N_RANGE	Différence entre le max et le min des Mean2 du rapport Harmonique/Bruit	$1 \le n \le N$
H2N_MEAN	Moyenne des Mean1 des rapports Harmonique/Bruit	

H2N_VAR	Variance des Mean1 des rapports Harmonique/Bruit	
Descripteurs Formantiques		
F2_F1_MIN	Minimum des différences entre les mean2 des fréquences des 1er et 2nd formants	$1 \le n \le N$
FI_FREQ_MEAN2	Mean2 de la fréquence du 1er formant	
FI_FREQ_MEAN1	Mean1 de la fréquence du 1er formant	
F2_FREQ_MEAN2	Mean2 de la fréquence du 2nd formant	
F2_FREQ_MEAN1	Mean1 de la fréquence du 2nd formant	
F3_FREQ_MEAN2	Mean2 de la fréquence du 3ème formant	
F3_FREQ_MEAN1	Mean1 de la fréquence du 3ème formant	
F2_RATIO_MEAN1	Fréquence du 2nd formant divisée par la différence entre les mean1 des fréquences des 1er et 2nd formants	$f2ratio_{Mean\,1} = \frac{f2_{Mean1}}{f2_{Mean\,1} - f1_{Mean\,1}} = \frac{QF.2b}{QF.0b}$
F2_RATIO_MEAN2	Fréquence du 2nd formant divisée par la différence entre les mean2 des fréquences des 1er et 2nd formants	$f2ratio_{Mean\,2} = \frac{f2_{Mean\,2}}{f2_{Mean\,2} - f1_{Meaan\,2}} = \frac{QF.2a}{QF.0a}$
F2_RATIO_MAX	Maximum des ratios précédents sur toute la phrase	$f2ratio_{\max} = \max_{nframes}\left(\frac{f2_i}{f2_i - f1_i}\right)$
F2_RATIO_RANGE	Différence entre le maximum et le minimum des ratios précédents sur toute la phrase	$f2ratio_{range} = \max_{nframes}\left(\frac{f2_i}{f2_i - f1_i}\right) - \min_{nframes}\left(\frac{f2_i}{f2_i - f1_i}\right)$
F1_BW_MEAN1	Moyenne des Mean1 des largeurs de bandes du 1er Formant	$bw_1 = \frac{\sum_{n=1}^{N} b_{1,n}}{N}$
F1_BW_MEAN2	Moyenne des Mean2 des largeurs de bandes du 1er Formant	
F1_BW_REG	Moyenne arithmétique des largeurs de bandes du 1er Formant calculée pour chaque trame d'une région voisée	
F2_BW_MEAN1	Moyenne des Mean1 des largeurs de bandes du 2nd Formant	$bw_2 = \frac{\sum_{n=1}^{N} b_{2,n}}{N}$

F2_BW_MEAN2	Moyenne des Mean2 des largeurs de bandes du 2ndFormant	
F2_BW_REG	Moyenne arithmétique des largeurs de bandes du 3ème Formant calculée pour chaque trame d'une région voisée	
F3_BW_MEAN1	Moyenne des Mean1 des largeurs de bandes du 3ème Formant	$bw_3 = \frac{\sum_{n=1}^{N} b_{3,n}}{N}$
F3_BW_MEAN2	Moyenne des Mean2 des largeurs de bandes du 3èmeFormant	
F3_BW_REG	Moyenne arithmétique des largeurs de bandes du 3ème Formant calculée pour chaque trame d'une région voisée	
F1_MAX_REG	Maximum de la fréquence du 1er formant pour chaque région voisée	
F2_MAX_REG	Maximum de la fréquence du 2nd formant pour chaque région voisée	
F3_MAX_REG	Maximum de la fréquence du 3ème formant pour chaque région voisée	
F1_RANGE_REG	Différence entre le max et le min de la fréquence du 1er formant pour chaque région voisée	$f_{1,range} = \max_{nframes}(f_{1,i}) - \min_{nframes}(f_{1,i})$
F2_RANGE_REG	Différence entre le max et le min de la fréquence du 2nd formant pour chaque région voisée	
F3_RANGE_REG	Différence entre le max et le min de la fréquence du 3ème formant pour chaque région voisée	
F1_VAR_REG	Variance de la fréquence du 1er formant pour chaque région voisée	
F2_VAR_REG	Variance de la fréquence du 2nd formant pour chaque région voisée	
F3_VAR_REG	Variance de la fréquence du 3ème formant pour chaque région voisée	
Descripteurs Bandes de fréquence		
ENER_BAND_1	Énergie mesurée dans la 1ère bande	0 Hz < f < f0
ENER_BAND_2	Énergie mesurée dans la 2ème bande	0 Hz < f < 1 KHz

ENER_BAND_3	Énergie mesurée dans la 3ème bande	2,5 KHz < f < 3,5 KHz
ENER_BAND_4	Énergie mesurée dans la 4ème bande	4 KHz < f < 5 KHz
ENER_BAND_RATIO	Énergie contenue dans les 4 bandes divisée par l'énergie totale	
ENER_RATE	Rapport entre les énergies contenues dans les régions voisées et l'énergie totale	$1 \le n \le N$
ENER_REL_REG	Énergie d'une région voisée divisée par l'énergie totale	$Ene\,\mathrm{Re}l_n = \frac{ene_n}{AbsEne}$
Descripteurs de la glotte		
OPEN_QUOTIENT	Différence entre les amplitudes des 1er et 2nd partiels	relatif au quotient d'ouverture de la glotte
OPEN_QUOTIENT_INVERSE	Différence entre les amplitudes des 1er et 2nd partiels après filtrage inverse (de la fonction de transfert du conduit vocal)	
SPECTRAL_TILT	Différence entre l'amplitude du 1er harmonique et l'amplitude du 2nd formant	Paramètre positif pour une voix soufflée et négatif pour une voix grinçante
SPECTRAL_TILT_INVERSE	Différence entre l'amplitude du 1er harmonique et l'amplitude du 2nd formant après filtrage inverse	

Définitions de descripteurs acoustiques issus de (Pereira, 2000)

Logiciels utilisés pour l'étude de la voix:

- **Praat** (University of Amsterdam) http://www.fon.hum.uva.nl/praat/
- **Wavesurfer** (KTH) http://www.speech.kth.se/wavesurfer/
- **Xspect** (IRCAM) http://mediatheque.ircam.fr/articles/textes/Rodet96b/
- **Talkapillar** (IRCAM) http://recherche.ircam.fr/equipes/analyse-synthese/concat/

Appendix B:

« A hybrid concatenative synthesis system
on the intersection of music and speech »
Acte du JIM 2005, présenté le 2 Juin 2005

A HYBRID CONCATENATIVE SYNTHESIS SYSTEM ON THE INTERSECTION OF MUSIC AND SPEECH

Grégory Beller, Diemo Schwarz, Thomas Hueber, Xavier Rodet
Ircam – Centre Pompidou,
Institut de Recherche et de Coordination Acoustique/Musique
1, place Igor Stravinsky
75004 Paris, France
{beller,schwarz,hueber,rodet}@ircam.fr

ABSTRACT

In this paper, we describe a concatenative synthesis system which was first designed for a realistic synthesis of melodic phrases. It has since been augmented to become an experimental TTS (Text-to-Speech) synthesizer. Today, it is able to realize hybrid synthesis involving speech segments and musical excerpts coming from any recording imported in its database. The system can also synthesize sentences with different voices, sentences with musical sounds, melodic phrases with speech segments and generate compositional material from specific intonation patterns using a prosodic pattern extractor.

1. INTRODUCTION

Musical concatenative synthesis consists in choosing the most appropriate sequence of sound units from a database and applying various modifications in order to build a desired melodic phrase. In previous work [12, 13, 14, 15] a concatenative musical sound synthesis system named CATERPILLAR has been elaborated. This work has been continued and an experimental TTS system, named TALKAPILLAR, is under construction. One of the aims of this system is to reconstruct the voice of a speaker, for instance a deceased eminent personality. TALKAPILLAR should pronounce texts as if they has been pronounced by the target specific speaker. The main difficulty consists in choosing the best speech segments, called units, to produce intelligible and natural speech with respect to the expressive characteristics of the speaker.

CATERPILLAR and TALKAPILLAR are based on the same architecture and thus allow to create hybrid synthetic phrases with speech units and any other sound units [3, 2].

After a quick overview of related work, this article explains the system and the successive steps of source sound analysis, database, and synthesis. Then it proposes several applications which could lead to new musical experiments.

2. RELATED WORK

2.1. Speech synthesis

Research in music synthesis is influenced by research in speech synthesis to which considerable efforts have been devoted. Concatenative unit selection speech synthesis from large databases, also called corpus based synthesis [5], is now used in many TTS systems for waveform generation [8]. Its introduction resulted in a considerable gain in quality of the synthesized speech over rule-based parametric synthesis systems, in terms of naturalness and intelligibility. Unit selection algorithms attempt to find the most appropriate speech unit in the database (corpus) by using linguistic features computed from the text to synthesize. Selected units can be of any length (non-uniform unit selection), from sub-phonemes to whole phrases, and are not limited to diphones or triphones. Although concatenative sound synthesis is quite similar to concatenative speech synthesis and shares many concepts and methods, both have different goals (see [15] for more details).

2.2. Singing Voice Synthesis

Concatenative singing voice synthesis occupies an intermediate position between speech and sound synthesis, whereas the used methods are most often close to speech synthesis [6]. There is a notable exception [7], where an automatically constituted large unit database is used. See [10] for an up-to-date overview of current research in singing voice synthesis, which is out of the scope of this article.

3. GENERAL OVERVIEW

All the processes involved, for music and speech analysis and hybrid synthesis, are presented in figure 1. They will be explained in more detail in the following sections.

4. SOUND SEGMENTATION

The first step is the segmentation of recorded files, speech or music, in variable length units. For this purpose, sound

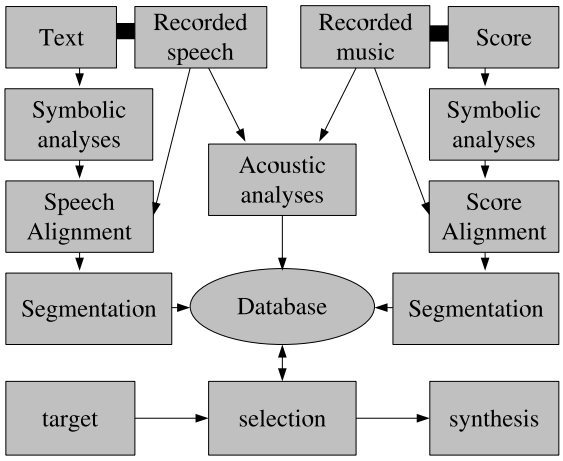


Figure 1. Architecture of the hybrid concatenative synthesis system.

files are first time-aligned with their symbolic representation (phonetic text or score) where unit frontiers are known. Then, units are imported into a POSTGRESQL database.

4.1. Score alignement

Score to audio alignment, or in brief score alignment, connects events in a score to corresponding points on the performance signal time axis [17]. A very similar task is known as score following, this term being reserved for the real-time case such as the one where a computer program is used to accompany a musician. Score alignment can thus use the whole score and the whole audio file if needed to perform the task, while score following specifies a maximum latency between an event in the audio stream and the decision to connect it to an event in the score. By using score information, score alignment permits to perform extensive audio indexing. It allows computing note time-onset, duration, loudness, pitch contour, descriptors and interpretation in general. Automatic score alignment has many applications among which we can mention: Audio segmentation into note samples for data base construction.

4.2. Speech alignement

A slightly different process is used to segment a speech signal. From the phonetic transcription of the sentence to align, a rudimentary synthesized sentence is built with diphones extracted from a small database where diphones frontiers have been placed by hand. Then the MFCC sequences of the two sentences are computed and aligned with a DTW algorithm. This provides the frontiers of phones and diphones in the sentence to align (see figure 2).

5. ACOUSTIC ANALYSIS

5.1. Musical Descriptors

We distinguish three types of descriptors: *Category descriptors* are boolean and express the membership of a

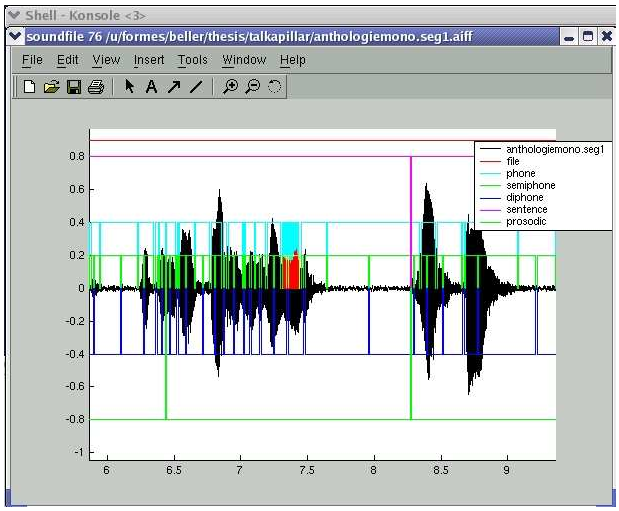


Figure 2. Example of speech segmentation due to alignment.

unit to a category or class and all its base classes in the hierarchy (e.g. violin → strings → instrument for the sound source hierarchy). *Static descriptors* take a constant value for a unit (e.g. Midi note number), and *dynamic descriptors* are analysis data evolving over the unit (e.g. fundamental frequency). The descriptors used are given in the following [11]. The perceptual salience of some of these descriptors depends on the sound base they are applied to.

Signal and Perceptual Descriptors
: Energy, fundamental frequency, zero crossing rate, loudness, sharpness, timbral width

Spectral Descriptors
: Spectral centroid, spectral tilt, spectral spread, spectral dissymmetry

Harmonic Descriptors
: Harmonic energy ratio, harmonic parity, tristimulus, harmonic deviation

Temporal Descriptors
: Attack and release time, ADSR envelope, center of gravity/antigravity

Source and Score Descriptors
: Instrument class and subclass, excitation style, Midi pitch, lyrics (text and phonemes), other score information

All of these, except the symbolic source and score descriptors, are expressed as vectors of *characteristic values* that describe the evolution of the descriptor over the unit:

- arithmetic and geometric mean, standard deviation
- minimum, maximum, and range slope, giving the rough direction of the descriptor movement, and curvature (from 2nd order polynomial approximation)
- value and curve slope at start and end of the unit (to calculate a concatenation cost)

- the temporal center of gravity/antigravity, giving the location of the most important elevation or depression in the descriptor curve and the first 4 order temporal moments
- the normalized Fourier spectrum of the descriptor in 5 bands, and the first 4 order moments of the spectrum. This reveals if the descriptor has rapid or slow movement, or if it oscillates.

5.2. Prosodic Features: YIN

Some acoustic features are extracted in order to build prosodic units of a specific speaker. Fundamental frequency is calculated with the YIN algorithm [4]. This also gives an energy and an descriptor of the aperiodicity for each frame. These data are averaged for each unit, sorted by phonetic classes and compared all together in order to build a distortion model relative to particular intonation strategies. We also compare the lasts of units, ones compared to the others per phonetic classes. This leads to an acoustic representation of prosodic units independent of the phonetic context.

6. SYMBOLIC ANALYSIS

The symbolic description of musical units is a midi score. The phonetic and syntactic description of texts is provided by the Euler program [1] issued from the European TTS project MBROLA [9]. This module analyzes a text and gives several symbolic representations such as a phonetic transcription, a grammatical analysis, a prediction of prosodic boundaries, etc. These descriptions are joined together in a MLC (Multi Layer Container) file (see figure 3). This file content and other symbolic descriptors corresponding to the relative places of the units are stored into an SDIF file (Sound Description Interchange Format, see [16]), which is the final description of segmental and supra-segmental units.

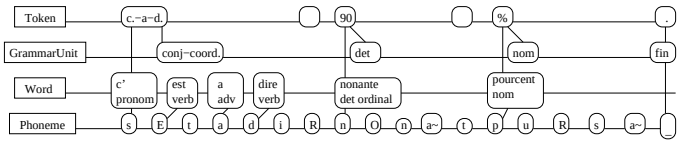


Figure 3. Example of an MLC structure [1] (from http://tcts.fpms.ac.be/synthesis/euler/doc/applicationdeveloper).

7. DATABASE

7.1. Database Management System

Since synthesis quality increases with the size of the sound database, an efficient database architecture is required. A relational DBMS (database management system) is used in this project. Although this results in a performance penalty for data access, the advantages in data handling prevail:

Data Independence
In a relational database, only the logical relations in the database schema are specified, not the physical organization of the data. It is accessed by the declarative query language SQL, specifying what to do, not how to do it, leading to unprecedented flexibility, scalability and openness to change.

Consistency
It is assured by atomic transactions. A transaction is a group of SQL statements which are either completely executed, or rolled back to the previous state of the database, if an error occurred. This means no intermediate inconsistent state of the database is ever visible. Other safeguards are the consistency checks according to the relations between database tables given by the schema, and the automatic reestablishment of referential integrity by cascading deletes (for example, deleting a sound file deletes all the units referring to it, and all their data). This is also an enormous advantage during development, because programming errors cannot corrupt the database.

Client–Server Architecture
Concurrent multi-user access over the network, locking, authentication, and fine-grained control over user's access permission are standard features of DBMS. In our system, this allows, for instance, to run multiple processes simultaneously in order to increase computation speed.

7.2. Database Interface

The database is clearly separated from the rest of the system by a database interface, dbi, (see figure 4), written in Matlab and procedural SQL. Therefore, the current DBMS can be replaced by another one, or other existing sound databases can be accessed, e.g. using the MPEG-7 indexing standard, or the results from the CUIDADO [18] or ECRINS [11] projects on sound content description. The database can be browsed with a graphical database explorer called dbx that allows users to visualize and play the units in the database. For low-level access, we wrote C-extensions for Matlab that send a query to the database and return the result in a matrix. For convenience, the database interface also centralizes access to external sound and data files. For the latter, SDIF is used for well-defined exchange of data with external programs (analysis, segmentation, synthesis). Finally the free, open source, relational DBMS POSTGRESQL, is used to reliably store thousands of sound and data files, tens of thousands of units, their interrelationship and descriptor data.

8. UNIT SELECTION

8.1. Features Involved

After the importing step, the database is filled with plenty of heterogeneous units: diphones, prosodic groups, dinotes

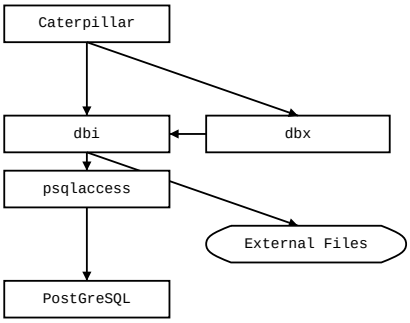


Figure 4. Architecture of the database interface.

(from the middle of a note to the middle of the next), attacks, sustains... Each of them is described by a set of features involving symbolic and acoustic analyses. These features are the basis for comparison between units and unit selection. Adequacy of units (target cost) and concatenation cost are computed as a weighted combination of the features. Minimization of the global cost of a unit selection is done by a Viterbi algorithm

8.2. Unit Selection Algorithm

The classical unit selection algorithm finds the sequence of database units u_i that best match the given synthesis target units t_τ using two cost functions: The *target cost* expresses the similarity of u_i to t_τ including a context of r units around the target. The *concatenation cost* predicts the quality of the concatenation of u_i with a preceding unit u_{i-1}. The optimal sequence of units is found by a Viterbi algorithm finding the best path through the network of database units.

8.3. Prosodic Unit Selection

The first step of TALKAPILLAR is to select supra-segmental units in the database. By a Viterbi algorithm applied on the symbolic representation of prosodic units, we access to the best prosodic sequence able to traduce the natural expressivity of the speaker. Then we add the acoustic parameters of the chosen prosodic units to the segmental units of the target in the aim of select the best sequence of segmental units that fit to a real prosody excerpt from the database.

8.4. Segmental Units Selection

The second step consists in selecting the segmental units according to their symbolic representation and to the acoustic representation derived from the selection of prosodic units coming from the previous step. A similar Viterbi selection algorithm is then applied to find the best sequence of segmental units that match the target string and the prosody selected.

9. SYNTHESIS

9.1. Concatenation

When a correct sequence of units has been selected, it just needs to be concatenated in order to build the desired phrase. This is the last step of the synthesis process and could be aimed at two different goals. Concatenative synthesis has been designed to preserve all sound details so as to improve the quality and the naturalness of the result. In the TALKAPILLAR TTS system, a first strategy is to not transform chosen units. They are concatenated with a slight cross fade at the junction and a simple period alignment to not produce clicks and other artefacts.

9.2. Transformation: PSOLA

Another strategy consists in transforming some of the selected units before concatenating them. For instance, some of the units (the voiced ones) are slightly time-stretched and pitch-transposed to best match the prosody selected beforehand. These transformations are accomplished with a PSOLA analysis-synthesis algorithm [9].

10. APPLICATIONS

10.1. TTS synthesis

The TALKAPILLAR synthesis system is not aimed at a classical TTS use as is the case of the majority of similar TTS systems. It is mainly designed to (re)produce a specific expressivity. It offers an excellent framework for an artistic purpose since the user has access to all the steps of the process, which permits a full control of the result.

10.2. Musical Synthesis

From a Midi score or a target sound file, one can, for instance, generate the most similar musical phrase with the sequence of units extracted from another sound file. This means that the system is able, for example, to synthesize a realistic violin phrase containing the same melody as a recorded voiced line. It simply finds trough its multiple representation of the sound units, the best sequence that matches the given target.

10.3. Hybrid Synthesis

An interesting aspect of the system is that the TTS synthesizer and the musical synthesizer share the same framework. Created on the same software architecture, these two synthesis systems joined together give a powerful tool for composers interested in interaction between music and speech. For instance, voiced parts of a sentence could be replaced by cello's sustained units respecting the prosody of the replaced speech segments. The flexibility of the selection process's parametrization (features involved, cost weights, etc.) sets the user free to create very innovative hybrid synthetic phrases.

10.4. Prosody Extraction

An option of the system permits to extract prosodic units out of the database. Exportation of the acoustic features (f0, energy, flow, etc.) and of the symbolic descriptors (grammatical structure, type of the final accent, etc.) into an SDIF file allows to exploit these supra-segmental groups in other musical software such as OpenMusic, Diphone, or MAX/MSP.

11. CONCLUSION

In this paper, we have described a new concatenative synthesis system able to deal with speech and other sound material. The different steps of the process have been presented so as to provide a global comprehension of the system. The combination of two applications, TTS and musical phrases synthesis, in the same framework augments the capacities for artistic applications. Some examples and a mailing list for discussions about concatenative synthesis in general can be found at:

http://recherche.ircam.fr/anasyn/concat
http://lists.ircam.fr/wws/info/concat

12. REFERENCES

[1] Michel Bagein, Thierry Dutoit, Nawfal Tounsi, Fabrice Malfrère, Alain Ruelle, and Dominique Wynsberghe. Le projet EULER, Vers une synthèse de parole générique et multilingue. *Traitement automatique des langues*, 42(1), 2001.

[2] Grégory Beller. La musicalité de la voix parlée. Maitrise de musique, Université Paris 8, Paris, France, 2005.

[3] Grégory Beller. Un synthétiseur vocal par sélection d'unités. Rapport de stage DEA ATIAM, Ircam – Centre Pompidou, Paris, France, 2004.

[4] Alain de Cheveigné and Hideki Kawahara. YIN, a Fundamental Frequency Estimator for Speech and Music. *Journal of the Acoustical Society of America (JASA)*, 111:1917–1930, 2002.

[5] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, Atlanta, GA, May 1996.

[6] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George. Concatenation-Based MIDI-to-Singing Voice Synthesis. In *103rd Meeting of the AES*. New York, 1997.

[7] Yoram Meron. *High Quality Singing Synthesis Using the Selection-based Synthesis Scheme*. PhD thesis, University of Tokyo, 1999.

[8] Romain Prudon and Christophe d'Alessandro. A selection/concatenation TTS synthesis system: Databases developement, system design, comparative evaluation. In *4th Speech Synthesis Workshop*, Pitlochry, Scotland, 2001.

[9] Prudon R., d'Alessandro C., and Boula de Mareüil. Prosody Synthesis by Unit Selection and Transplantation on Diphones. In *IEEE 2002 Workshop on speech synthesis*, Santa Monica, USA, 2002.

[10] Xavier Rodet. Synthesis and processing of the singing voice. In *Proceedings of the 1stIEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA)*, Leuven, Belgium, November 2002.

[11] Xavier Rodet and Patrice Tisserand. ECRINS: Calcul des descripteurs bas niveaux. Technical report, Ircam – Centre Pompidou, Paris, France, October 2001.

[12] Diemo Schwarz. A System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 97–102, Verona, Italy, December 2000.

[13] Diemo Schwarz. New Developments in Data-Driven Concatenative Sound Synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 443–446, Singapore, October 2003.

[14] Diemo Schwarz. The CATERPILLAR System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 135–140, London, UK, September 2003.

[15] Diemo Schwarz. *Data-Driven Concatenative Sound Synthesis*. Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris, 2004.

[16] Diemo Schwarz and Matthew Wright. Extensions and Applications of the SDIF Sound Description Interchange Format. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 481–484, Berlin, Germany, August 2000.

[17] Ferréol Soulez, Xavier Rodet, and Diemo Schwarz. Improving Polyphonic and Poly-Instrumental Music to Score Alignment. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, Maryland, USA, October 2003.

[18] Hugues Vinet, Perfecto Herrera, and François Pachet. The Cuidado Project: New Applications Based on Audio and Music Content Description. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 450–453, Gothenburg, Sweden, September 2002.

Appendix C:

« Multi-Speaker ESS with Only One Expressive Database »

Article soumis à IEEE “special issue on ESS”, le 1er Juin 2005

Multi-Speaker ESS with Only One Expressive Database

Grégory Beller, Fernando Villavicencio, Thomas Hueber, Diemo Schwarz
Ircam, Institut de Recherche et de Coordination Acoustique/Musique
1, place Igor Stravinsky
75004 Paris, France
{beller,villavicencio,hueber,schwarz}@ircam.fr

***Abstract*— In this paper we describe a new expressive speech synthesizer. Mainly based on a Text-To-Speech (TTS) system, it gives the opportunity to a user to specify a sentence and an affect and produces automatically an affected utterance. The result is refined by transformations based on a phase vocoder technology. Those transformations are induced by a supra-segmental selection step. The supra-segments called extended prosodic units contains evolutions of acoustic parameters influenced by emotions. The way they are defined makes the overall system speaker-independent since it is assumed that identity voice conversion is equivalent as expressive state conversion in an intra-speaker way. Thus, the system synthesizes and transforms neutral utterances into affected ones without the need of prerecorded affected diphones. The system is now working for several emotions in french and is used for an artistic purpose dealing with multimedia and cinema.**

I. Introduction

Current concatenative synthesis methods have significantly improved the naturalness of synthetic speech. Concatenative synthesis consists in choosing the most appropriate sequence of speech units from a database and applying various modifications in order to build a desired sentence. In previous work [34], [35], [36] a musical concatenative system named CATERPILLAR has been elaborated. This work has been continued and a Text-to-Speech (TTS) system, named TALKAPILLAR[2], is born. One of the aims of this system is to reconstruct the voice of a speaker, for instance a deceased eminent personality. TALKAPILLAR should pronounce texts as if they has been pronounced by the target specific speaker. The main difficulty consists in choosing the best speech segments, called units, to produce intelligible and natural speech with respect to the expressive characteristics of the speaker. This aim has been enlarged to become a multi-voice ESS.

We have recorded a french actor to build an emotional speech database. Then we have constructed extended prosodic models of the ways he has conveyed emotions by acoustical analysis of the speech signal. The normalization of the evolutions of acoustic parameters that occur in expressive utterances permits to use these prosodic strategies for another speaker. The fact is that we assume that changing voice identity from a speaker A to a speaker B is the same processus as changing expressive state for one speaker. Thus, we define the characterization of the identity of a speaker and of its expressive state by the same prosodic model. That's why the expressive state of another speaker could be extrapolated by a simple composition of its identity's characteristics and the characteristics of the expressive state we want to synthesize. The use of those extended prosodic models as targets in the selection step and in the transformation step leads us to an expressive speech synthesis. The system allows the user to transform the synthesis with its own expressibility.

After a quick overview of related work, this article explains the system and the successive steps of source analysis, database and synthesis. Then it proposes several applications.

II. Related Work

A. Speech synthesis

Considerable efforts have been devoted to research in speech synthesis. Concatenative unit selection speech synthesis from large databases, also called corpus based synthesis [15], is now used in many TTS systems for waveform generation [27]. Its introduction resulted in a considerable gain in quality of the synthesized speech over rule-based parametric synthesis systems, in terms of naturalness and intelligibility. Unit selection algorithms attempt to find the most appropriate speech unit in the database (corpus) by using linguistic features computed from the text to synthesize. Selected units can be of any length (non-uniform unit selection), from sub-phonemes to whole phrases, and are not limited to diphones or triphones.

B. Expressive Speech synthesis

Recently, the development of corpus based methods and the increasing size of databases widens TTS systems towards ESS. [5] recorded several speech corpora pronounced under different emotions and used classic TTS methods on these separated databases to provide ESS. [18] made as well three databases (anger, joy and sadness) and switched between these emotional corpora to synthesize the desired corresponding affected utterance. An attempt to group the different corpora without formal separations into a same speech database have been made by [7]. As many other TTS systems, [14] build expressive prosody models from the data and then use them to find the best segmental units sequence. [41] uses a HMM-training to produce ESS with a limited speech database. To see an alternative of corpus based methods for ESS, see [8].

C. Singing Voice and Music Synthesis

Concatenative singing voice synthesis occupies an intermediate position between speech and sound synthesis, whereas the used methods are most often close to speech synthesis [19]. There is a notable exception [21], where an automatically constituted large unit database is used. Hybrid concatenative synthesis on the intersection of music and speech have been tried with our concatenative synthesizer [4]. See [30] for an up-to-date overview of current research in singing voice synthesis, which is out of the scope of this article.

III. General Overview

All the processes involved are presented and summarized in figure 1. They will be explained in more detail in the following sections.

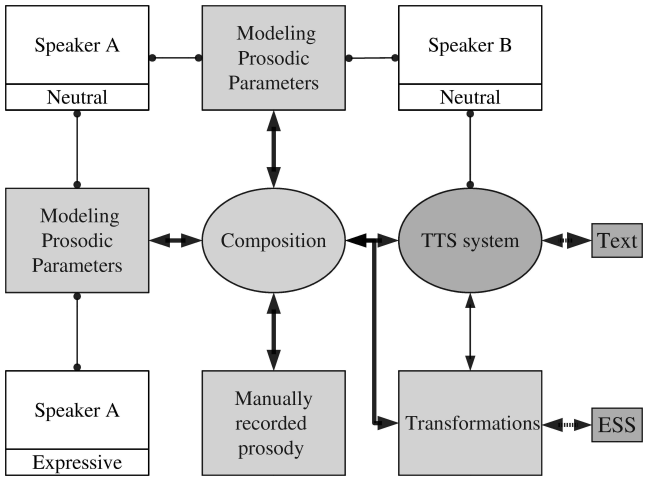


Fig. 1. Architecture of the Expressive Speech Synthesizer.

Building expressive speech synthesis needs to manage with two information levels conveyed by the speech. On one hand, cognitive parts of emotion's expression are conveyed trough prosodic clues. For instance, it is often related that happiness is demonstrated by an acceleration of the speech rate and an increase of the mean of the fundamental frequency. On the other hand, physiological parts of the emotion's expression [32] are demonstrated by uncontrolled deviations of the voice quality such as jitter. A good example is the increased brightness in a angry utterance. A good expressive speech synthesizer should thus involve processes of these two information levels. This is the aim of our synthesizer.

In addition, we try to build a speaker-independent system. Almost all studies on ESS uses neutral speech as reference. They compare expressive utterances to neutral ones in order to construct high level descriptors related to global prosodic deviations such as a f0-mean rise, a speech rate decrease or a jitter variation. The TTS context used here involve a supra-segmental selection step [20]. This step provides usually a prosodic context which leads to a better segmental selection as we applied real acoustic parameters to the target to find the best segmental units. By comparing high level features of these supra-segmental units to a fixed reference (computed thanks to neutral utterances), we can transform prosodic groups of a speaker A to prosodics groups of a speaker B. That means that we use neutral utterances comparisons between two speakers to change the identity of prosodic groups. For instance, we apply a multiplication with a factor-k to the f0-range of the intonation curves of a man to fit the f0-range of the intonation curves of a woman where k has been calculated by dividing mean-ranges of woman's neutral utterance by man's ones.

The main hypothesis of this paper is that such transformations used to convert identity's prosodic groups could be use for converting expressive prosodic deviations for one speaker. That means that identity's conversion is similar to expressive state conversion. Assuming this hypothesis, we can compose those transformations (see figure 2) to build expressive utterances from only neutral utterances of a speaker B, thanks to a scaling process of prosodic differences between neutral and expressive utterances of a speaker A. By applying such global transformation to expressive prosodic groups, we can use prosodic groups of a speaker A to generate prosodics groups for a speaker B (see figure 2).

It is a great advance in ESS since it needs only one expressive database made by one speaker to synthesize many expressive voices whom only a neutral database recording is needed. Furthermore, the application give the opportunity to "improve" prosody after the selection by recording a new one and to acoustically transform the resulting synthesized sentence thanks to phase vocoder algorithm [6].

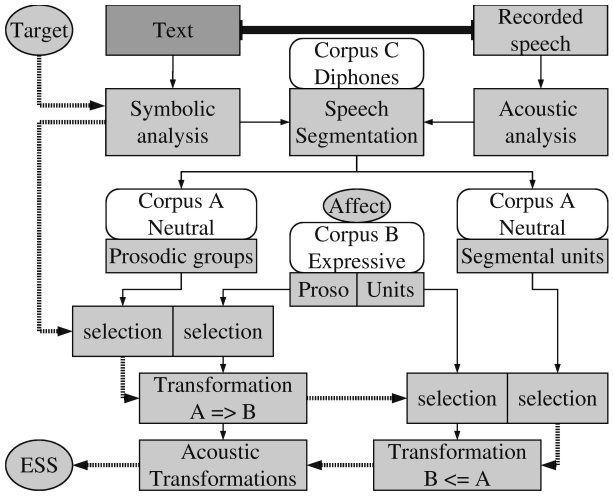


Fig. 2. More detailed architecture of the Expressive Speech Synthesizer. Corpora A, B and C are described in section IV-A.

IV. Database

A. Database Content

For this experiments, we built three databases. The first one is a diphone corpus pronounced by a speaker A and used for the alignment reference. The second is a large corpus of neutral

speech from a speaker A. The third corpus is composed by neutral and emotional utterances pronounced by a speaker B.

1) Corpus A: Neutral speech: The Corpus A contains an approximately two hours of neutral speech. The speaker A is a french man. He has been recorded reading a text with a high quality equipment in an anechoic chamber.

2) Corpus B: Expressive speech: The speaker B is a french actor. About two hours of emotional speech has been also recorded in an anechoic chamber. The corpus is composed by a set of 26 short sentences. Each was pronounced three times under different emotions (neutral, neutral question, angry, happy, sadness, bore, disgust, indignation, positive and negative surprises). The three occurrences correspond to a change of the activation level (low, middle, heavy). After post-processing, we kept 539 sentences for the analysis.

3) Corpus C: Neutral diphones: This database is a collection of diphones. The speaker A has pronounced all the french diphones as ("papap"). This set has been segmented manually to provide a referential time division of diphones. Each french diphone has been pronounced one time in a neutral mono-tonal expressivity.

B. Database Management System

Since synthesis quality increases with the size of the sound database, an efficient database architecture is required. A relational DBMS (database management system) is used in this project. Although this results in a performance penalty for data access, the advantages in data handling prevail:

Data Independence
: In a relational database, only the logical relations in the database schema are specified, not the physical organization of the data. It is accessed by the declarative query language SQL, specifying what to do, not how to do it, leading to unprecedented flexibility, scalability and openness to change.

Consistency
: It is assured by atomic transactions. A transaction is a group of SQL statements which are either completely executed, or rolled back to the previous state of the database, if an error occurred. This means no intermediate inconsistent state of the database is ever visible. Other safeguards are the consistency checks according to the relations between database tables given by the schema, and the automatic re-establishment of referential integrity by cascading deletes (for example, deleting a sound file deletes all the units referring to it, and all their data). This is also an enormous advantage during development, because programming errors cannot corrupt the database.

Client–Server Architecture
: Concurrent multi-user access over the network, locking, authentication, and fine-grained control over user's access permission are standard features of DBMS. In our system, this allows, for instance, to run multiple processes simultaneously in order to increase computation speed.

C. Database Interface

The database is clearly separated from the rest of the system by a database interface, dbi, (see figure 3), written in Matlab and procedural SQL. Therefore, the current DBMS can be replaced by another one, or other existing speech databases can be accessed, e.g. using the MPEG-7 indexing standard, or the results from the CUIDADO [24] or ECRINS [23] projects on sound content description. The database can be browsed with a graphical database explorer called dbx that allows users to visualize and play the units in the database. For low-level access, we wrote C-extensions for Matlab that send a query to the database and return the result in a matrix. For convenience, the database interface also centralizes access to external sound and data files. For the latter, SDIF [37] is used for well-defined exchange of data with external programs (analysis, segmentation, synthesis). Finally the free, open source, relational DBMS POSTGRESQL, is used to reliably store thousands of sound and data files, tens of thousands of units, their interrelationship and descriptor data.

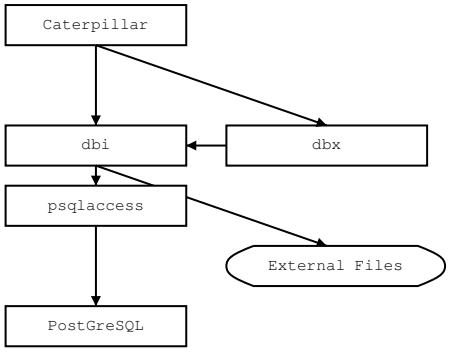


Fig. 3. Architecture of the database interface.

V. SPEECH SEGMENTATION BY ALIGNMENT

The first step is the segmentation of recorded files, in variable length units. For this purpose, sound files are first time-aligned with their symbolic representation (phonetic text) where unit frontiers are known. Then, units are imported into the POSTGRESQL database. Speech alignment connects events in a text to corresponding points on the speech signal time axis [38]. A very similar task is known as speech recognition, this term being reserved for the real-time case. Speech alignment can thus use the whole text and the whole audio file if needed to perform the task, while speech recognition specifies a maximum latency between an event in the audio stream and the decision to connect it to an event in the text. By using text information, speech alignment permits to perform extensive audio indexing. It allows computing phonetic time-onset, duration, loudness, pitch contour, high level descriptors and expressivity in general. Automatic speech alignment has many applications among which we can mention: Audio segmentation into phone samples for data base construction. Thus the process is used to segment a speech signal. From the phonetic transcription of the sentence to align, a rudimentary

synthesized sentence is built with diphones extracted from the small database called corpus A in section IV-A where diphones frontiers have been placed by hand. Then the MFCC sequences of the two sentences are computed and aligned with a DTW algorithm. This provides the frontiers of phones and diphones in the sentence to align (see figure 4).

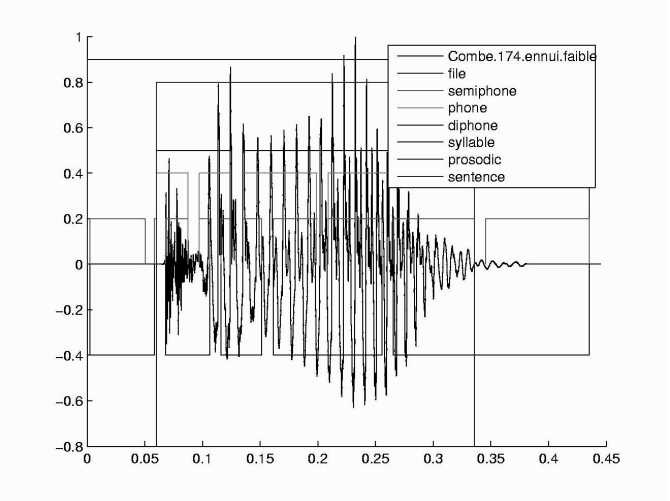


Fig. 4. Example of speech segmentation due to alignment.

VI. Symbolic Analysis

The phonetic and syntactic description of texts is provided by the EULERprogram [1] issued from the European TTS project MBROLA [28]. This module analyzes a text and gives several symbolic representations such as a phonetic transcription, a grammatical analysis, a prediction of prosodic boundaries, etc. These descriptions are joined together in a MLC (Multi Layer Container) file (see figure 5). This file content and other symbolic descriptors corresponding to the relative places of the units from one to each others are stored into an SDIF file (Sound Description Interchange Format, see [37]). Those files contain symbolic descriptions of segmental and supra-segmental units.

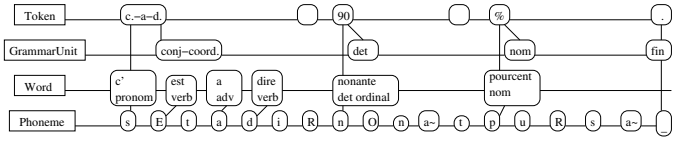


Fig. 5. Example of an MLC structure [1] (from http://tcts.fpms.ac.be/synthesis/euler/doc/applicationdeveloper).

VII. Descriptors

We distinguish three types of descriptors: *Category descriptors* are boolean and express the membership of a unit to a category or class and all its base classes in the hierarchy (e.g. speaker → actor → male for the sound source hierarchy). *Static descriptors* take a constant value for a unit (e.g. phoneme). *dynamic descriptors* are analysis data evolving over the unit (e.g. fundamental frequency). The descriptors used are given in the following sections. The perceptual salience of some of these descriptors depends on the sound base they are applied to.

Signal and Perceptual Descriptors
: Energy, fundamental frequency, jitter, shimmer

Spectral Descriptors
: First four formants frequencies, bandwidths, amplitudes and their derivatives.

Harmonic Descriptors
: Harmonic to noise ratio also called aperiodicity and drop-off energy

Temporal Descriptors
: Duration, speech rate

Source and Symbolic Descriptors
: Speaker, emotional state, text, phoneme, other contextual information...

All of these, except the symbolic source and Symbolic descriptors, are expressed as vectors of *characteristic values* that describe the evolution of the descriptor over the unit:

- arithmetic and geometric mean, standard deviation
- minimum, maximum, and range slope, giving the rough direction of the descriptor movement, and curvature (from 2nd order polynomial approximation)
- value and curve slope at start and end of the unit (to calculate a concatenation cost)
- the temporal center of gravity/anti-gravity, giving the location of the most important elevation or depression in the descriptor curve and the first 4 order temporal moments
- the normalized Fourier spectrum of the descriptor in 5 bands, and the first 4 order moments of the spectrum. This reveals if the descriptor has rapid or slow movement, or if it oscillates.

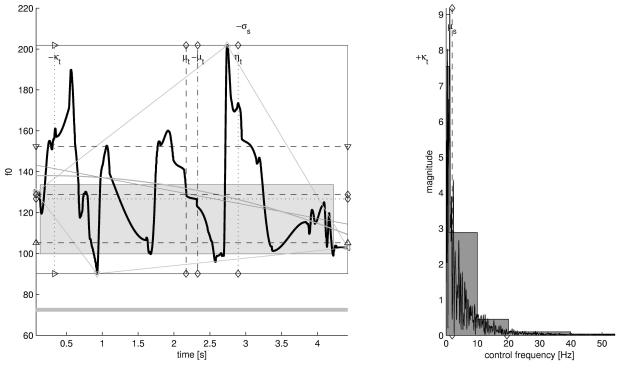


Fig. 6. Example of *characteristic values* of the fundamental frequency computed over all an angry utterance.

VIII. Voice characterization

The characterization of the identity of a speaker or of its expressive state use global prosodic descriptors. Mean, standard deviation, range, max and min of the *characteristic values*

computed for the fundamental frequency, the speech rate, the energy and the voice quality over all or several utterances give a parametrization of the speaker state and its identity. These high level descriptors are built with the same *characteristic values* as those computed for the description of segmental units trough we construct prosodic curves. The own difference is the temporal horizon for the computation of those values. A mean-f0 value over all neutral utterances characterizes the vocal register of a speaker whereas a mean-f0 value over a single vowel characterizes part of a prosodic movement. Therefore the mean-f0 values computed over segmental units and concatenated give a representation of the prosody as dynamic movements. The next step is to normalize those segmental curves by high-level features such as mean-f0 computed over all neutral utterances. As we represent dynamic deviations relative to the prosodic movements by normalized curves (e.g. f0, energy, speech rate and voice quality evolutions normalized by their contextual mean values), we can transplant expressive deviations from a speaker A to a speaker B with respect of his proper voice identity characterization. The generation of affected supra-segmental units is then just a scaling of those normalized curves by factors derived from high-level features describing the identity of a speaker.

IX. Acoustic Features

Some acoustic features are extracted in order to build prosodic units. They have been elaborated with the aim that their descriptions are non-dependent of the phonetic context. Thus, they are as much as possible only relatives to the ways a speaker pronounce and not to what he says. Following descriptors are dynamics and all the *characteristic values* presented on section VII are computed for each segmented units.

A. Fundamental frequency

Many papers related the fact that pitch curves also called intonation contour are the most prominent perceptual cue for expressivity. Fundamental frequency is calculated by an autocorrelation method with the YIN algorithm [10]. This algorithm also gives the energy and the Harmonic to noise ratio (also called aperiodicity) of the signal for each computed frame. In order to avoid local estimation errors like octave-jumps, the curve is median filtered. By thresholding the aperiodicity curve, we just keep f0-estimations on voiced part of the signal. Then the f0 of unvoiced parts of signal are defined by interpolating the values of voiced parts. Giving a f0 value to unvoiced parts is not common but offers a contextual information and draws a continuous curve of f0 called intonation contour. This curve is normalized by the mean of its overall f0-values to provide a speaker-independent evolution curve since people don't have the same mean f0-value. The jitter is computed here thanks to the normalized Fourier spectrum of the f0 in 5 bands, by the examination of the first 4 order moments of this spectrum.

B. Speech Rate

Contrary to the well defined and consensual definition of fundamental frequency, the speech rate (also called rhythm or speech flow) of the speech is hard to define. Since it really depends on the language, there is not any consensus on its definition [44]. [13] defines it as an overall movement of the utterance. It takes account of silencious pauses, breaks, accelerations and decelerations. He distinguishes the articulatory rate conveyed by syllables and the speech rate counted by words. For instance, the mean articulatory rate is 5,3 syllables per second in English whereas the mean speech rate is about 200 words per minute. A syllabic rate language like french constructs its rhythm not on a word accent, but on the recurrence of the syllables which tend to have the same durations. Thus, the normalization of speech rate is often made in an intra-syllabic way [22]. This hypothesis of a constant syllabic frequency has been reinforced in a study of production [42]. Therefore we see trough this hypothesis that if a syllable has a greater duration than the others, it is for a prosodic issue and not for a phonetic reason. As it seems that a consensus on the reality of the psycho-rhythmic role of the syllable in french has been established [44], and that the speech rate is merely related to syllable as to phones [31], we build speech rate curves by examining the durations of syllables (see figure 7). Thanks to the segmentation and the alignment with the symbolic syllable's boundaries given by EULER, the speech rate curve is designed by the durations of syllables. A deceleration corresponds to a rising of the curve and an acceleration is represented by a falling of the curve. The normalized speech rate curve is the value of the speech rate curve divided by the mean of those duration values over the utterance. This curve represents a better perceptual reality since we don't think about absolute values of syllable's durations when we speak but merely about accelerations and decelerations. Furthermore the normalization permits the speaker-independence since people don't have the same mean speech rate.

figure 7 shows several informations: The curve presents local maxima corresponding to decelerations. The syllables are coded in XSampa. The framed syllables are the ones considered as accentuated by the text-to-prosody generator of EULER. For this example, EULER gives a good prediction of the prosody pronounced by the actor although it has been designed for neutral speech and not for expressive utterances.

The main advantage of this description of the speech rate is its instantness. Many studies [9] [25], on expressive speech use global clues over all the utterance to describe the speech rate even if it is often in syllable per seconds.

C. Energy

The energy curve given by the YIN algorithm is extremely dependent of the phonetic context. Thus, it does not traduce a perceptual energetic emphasis at all. To provide such interesting perceptual clue, we construct energy curves after having imported all the units in the database by calculating statistics

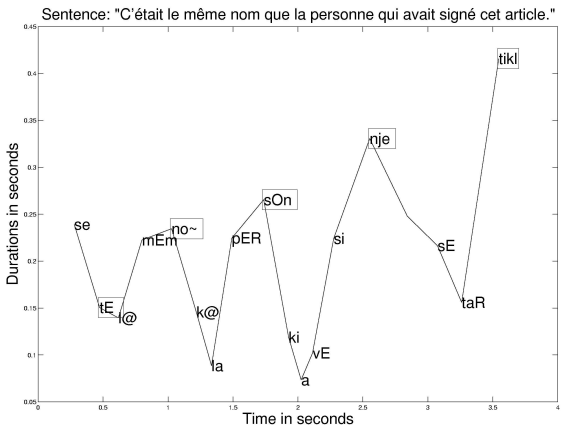


Fig. 7. Durations of syllables of a french sentence pronounced with happiness: "It was the same name as the person who had signed up this paper."

over phoneme's classes. First, frame-computed energy is averaged over the duration of each unit. Then wee group those means by phonetic classes and compared all together in order to build a distortion model relative to particular intonation strategies. We also compare the durations of units and their means of f0, ones compared to the others per phonetic classes. This leads to an acoustic representation of prosodic units independent of the phonetic context which aims to traduce the perceptual prosodic clues. The shimmer is estimated here thanks to the normalized Fourier spectrum of the energy in 5 bands, by the examination of the first 4 order moments of this spectrum.

D. Voice Quality

Like the energy, formants places are very dependent of the phonetic context. Thanks to the organization by phonemes, we access the deviations of formant's places for each phoneme pronounced under an affect. These normalized deviations combined with those of a speaker's identity changes give the opportunity to predict the formant's place for a phone pronounced by another speaker under an affect he has never pronounced.

Formants trajectories are estimated by the first maxima of the spectral envelop which is given by the true envelop algorithm. True envelop [16] is a very robust technique "rediscovered" and reported in [29] as a solution for the systematically errors presented in standard techniques as LPC or discrete cepstrum for spectral envelop estimation. Improved spectral envelop estimation techniques as Discrete LPC [12] or true envelop [16] outperforms the results obtained by the standard LPC analysis method. In this work we use a true envelop version of the speech spectrum to improve the spectral matching of the LPC modeling of voiced speech.

The drop-off energy feature is well described in [11]. It is the spectral tilt over 1000 Hz and is estimated by a mean-square distance minimization.

X. Unit Selection

A. Features Involved

After the importing step, the database is filled with plenty of heterogeneous units: semiphones, phones, diphones, syllables, prosodic groups... Each of them is described by a set of features involving symbolic and acoustic analysis. These features are the basis for comparison between units and unit selection. Adequacy of units (target cost) and concatenation cost are computed as a weighted combination of the features. Minimization of the global cost of a unit selection is done by a Viterbi algorithm

B. Unit Selection Algorithm

The classical unit selection algorithm finds the sequence of database units u_i that best match the given synthesis target units t_τ using two cost functions: The *target cost* expresses the similarity of u_i to t_τ including a context of r units around the target. The *concatenation cost* predicts the quality of the concatenation of u_i with a preceding unit u_{i-1}. The optimal sequence of units is found by a Viterbi algorithm finding the best path through the network of database units.

C. Prosodic Unit Selection

The first step of TALKAPILLAR is to select supra-segmental units in a database. By a Viterbi algorithm applied on the symbolic representation of prosodic units, we access to the best prosodic sequence able to traduce the a special expressivity of a speaker. By choosing those prosodic groups in an expressive database, we can have a representation of a real evolution of acoustic parameters for a chosen affect. After having been rescaled by a speaker-dependent identity factor, we add these acoustic parameters of the chosen and transformed prosodics units to the segmental units of the target in the aim of select the best sequence of segmental units that fit to a real prosody excerpt from the database.

D. Segmental Units Selection

The second step consists in selecting the segmental units according to their symbolic representation and to the acoustic representation derived from the selection of prosodic units coming from the previous step. A similar Viterbi selection algorithm is then applied to find the best sequence of segmental units that match the target string and the prosody selected.

XI. Synthesis

A. Concatenation

When a correct sequence of units has been selected, it just needs to be concatenated in order to build the desired phrase. This is the last step of the synthesis process and could be aimed at two different goals. Concatenative synthesis has been designed to preserve all sound details so as to improve the quality and the naturalness of the result. In the TALKAPILLAR TTS system, a first strategy is to not transform chosen units. They are concatenated with a slight cross fade at the junction and a simple period alignment to not produce clicks and other artifacts.

B. Acoustic Transformations

Another strategy consists in transforming some of the selected units before concatenating them. For instance, some of the units (the voiced ones) are slightly pitch-transposed to best match the prosody selected beforehand. They could also be time-stretched if the selected diphones length are to short compared to the desired speech rate. One can record the sentence with a different prosody and import it immediately in the database so as to provide new prosodic groups and then force the synthesis to follow its own expressivity. These transformations are accomplished with a phase vocoder based on a speech algorithm [6].

XII. Listening Test

Psycho-acoustic tests have been elaborated for several purposes. The first is naturally the verification of the expressive database. Some acted utterances could lead to a misunderstood affect. Thus we should test if people generally agrees with the commitment of the actor upon its own performance. After this validation step, we'll keep only valid matches for comparison. The second aim of the perceptive test is to watch over whether or not our synthesized utterances are expressive. Therefore part of the test are synthesized utterances. People are invited to classify the actor utterances mixed to the synthesized one. There are more emotional categories than in data to avoid a bias coming from the constraint of the selection [33]. Actor/Synthesizer's discrimination is also analyzed by a third measure. It correspond to the third goal of this study. By examining the correlation between the response-time and the distinction between actor's performance and synthesized ones, we expect a larger response-time when sentence are synthesized. The results of this psycho-acoustic test will be available in the late of 2005 since the system is still in fresh creation.

XIII. Applications

Interestingly, ESS is widely expected for its artistic purposes. Many composers and directors want to work with vocal expressivity. Furthermore, several computer-game's creators wish to add expressive voices to characters in non-predicted scenario. That aim naturally involves a TTS system and requires many voices as there are often a lot of characters in a computer-game. Thus, the solution of one general TTS module followed by multiple identity voice conversions seems to be the best way to reduce the processing time of the synthesis.

A. TTS synthesis

The TALKAPILLAR synthesis system is not aimed at a classical TTS use as is the case of the majority of similar TTS systems. It is mainly designed to (re)produce a specific expressivity. It offers an excellent framework for an artistic purpose since the user has access to all the steps of the process, which permits a full control of the result. A friendly user interface permits to avoid bad diphones (black list) and to transform the results with its own speech rate or f0 contour thanks to SVP. That's why interests for multimedia and cinema are shown and illustrated by the use of the system by a french dubbing company.

B. Hybrid synthesis

An interesting aspect of the system is that the TTS synthesizer TALKAPILLARand the musical synthesizer CATERPILLARshare the same framework. Created on the same software architecture, these two synthesis systems joined together give a powerful tool for composers interested in interaction between music and speech [4]. For instance, voiced parts of a sentence could be replaced by cello's sustained units respecting the prosody of the replaced speech segments. The flexibility of the selection process's parametrization (features involved, cost weights, etc.) sets the user free to create very innovative hybrid synthetic phrases. Experiments have been made in creating hybrid synthetic phrases with speech units and any other sound units [3].

C. Prosody Extraction

An option of the system permits to extract prosodic units out of the database. Exportation of the acoustic features (f0, energy, Speech rate, etc.) and of the symbolic descriptors (grammatical structure, type of the final accent, ...) into a SDIF file allows to exploit these informations in other frameworks.

XIV. Conclusion and Future Works

In this paper, we have described a new concatenative synthesis system able to synthesize expressive speech. The different steps of the process have been presented so as to provide a global comprehension of the system. The transplantation of expressive prosody to another speaker augments the capacities of ESS for artistic applications. Even if the system is already in use for artistic purposes, it has to be validated by measurements of the recognition rate during listening test to be compared to actual systems. Such results will appear before the end of the year 2005. Some examples can be yet downloaded from the following address: http://recherche.ircam.fr/equipes/analyse-synthese/concat

Many papers related the fact that prosodic parameters are not sufficient to provide a good ESS. Emotional states are linked to physiological changes that perturb or modify the voice quality. Like the perturbation of f0 called jitter, various uncontrolled spectral changes occur during emotional utterances. That's why we begin to include voice morphing methods and technology in our TTS framework. Voice conversion is usually defined as the adaptation of the characteristics of a source speaker's voice to those of a target speaker. Although the term is commonly interchanged by "voice morphing", they have been distinguished separately [26]. In speech morphing the source and target signals should be similar (same utterances, similar prosody and naturally time-aligned) to become reasonably aligned and interpolated for achieving new signals. On the other hand, voice conversion relies on the learning of source-target relationships from a number of (not necessary similar) utterances from two speakers, which are then used to

convert unseen signals of the source speaker towards the target speaker [26]. By using voice morphing techniques on isolated phonetic class, we aim to learn the transformations needed to convert speaker's identity and to convert from a neutral state to an expressive one. As a typical spectral conversion method, a mapping algorithm using a Gaussian Mixture Model (GMM) has been proposed originally by [39] and followed by several works [17], [43], [40]. Future works will be directed to a better treatment of voice quality using a mapping between two speakers and a mapping between two expressive states within one speaker. Then the proposed works will be to transform each neutral phones in regards of the combination of the two mappings.

ACKNOWLEDGMENTS

The authors would like to thanks the two speakers Jacques Combes and Xavier Rodet involved in this study for their performances.

REFERENCES

[1] Michel Bagein, Thierry Dutoit, Nawfal Tounsi, Fabrice Malfrère, Alain Ruelle, and Dominique Wynsberghe. Le projet EULER, Vers une synthèse de parole générique et multilingue. *Traitement automatique des langues*, 42(1), 2001.

[2] G. Beller. un synthétiseur vocal par sélection d'unités. rapport de stage, IRCAM, 2004.

[3] Grégory Beller. La musicalité de la voix parlée. Maitrise de musique, Université Paris 8, Paris, 2005.

[4] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet. A hybrid concatenative synthesis system on the intersection of music and speech. In *JIM*, IRCAM, 2005. Analysis/synthesis team.

[5] A.W. Black. Unit selection and emotional speech. *Eurospeech*, 2003.

[6] Niels Bogaards, Axel Roebel, and Xavier Rodet. Sound analysis and processing with audiosculpt 2. In *International Computer Music Conference (ICMC)*, Miami, USA, Novembre 2004.

[7] M. Bulut, S. Shrikanth, S.S. Narayanan, and A. K. Syrdal. Expressive speech synthesis using a concatenative synthesizer. In *ICSLP*, ATT Labs-Research, Florham Park, NJ, 2002.

[8] F. Burkhardt and W.F. Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Workshop on Speech and Emotion, Northern Ireland*, pages 151–156, 2000.

[9] S.-J. Chung. *L'expression et la perception de l'émotion extraite de la parole spontannée: évidences du coréen et de l'anglais*. phonétique, Université PARIS III - Sorbonne Nouvelle: Institut de Linguistique et Phonétique Générales et Appliquées, Paris, 2000. sous la direction de J. Vaissière.

[10] Alain de Cheveigné and Hideki Kawahara. YIN, a Fundamental Frequency Estimator for Speech and Music. *Journal of the Acoustical Society of America (JASA)*, 111:1917–1930, 2002.

[11] C. Drioli, G. Tisato, P. Cosi, and F. Tesser. Emotions and voice quality: Experiments with sinsoidal modeling. In *VOQUAL'03*, pages 127–132, Laboratory of Phonetics and Dialectology, ISTC-CNR, Institute of Cognitive Sciences and Technology, Italy, August 2003. drioli@csrf.pd.cnr.it.

[12] A. El-Jaroudi and J. Makhoul. Discrete all-pole modelling. *IEEE Transactions on signal processing*, 32:411–423, 1991.

[13] A. Galarneau, P. Tremblay, and P. Martin. dictionnaire de la parole. http://www.lli.ulaval.ca/labo2256/lexique/dico.html. Laboratoire de Phonétique et Phonologie de l'Université Laval à Québec.

[14] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli. The ibm expressive speech synthesis system. In *International Conference on Spoken Language Processing (ICSLP)*, IBM, October 2004.

[15] Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, Atlanta, GA, May 1996.

[16] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP*, pages 93–96, 1983.

[17] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *ICASSP*, pages 285–288, 1998.

[18] A. Lida, N. Campbell, F. Higuchi, and M. Yasumura. A corpus-based speech synthesis system with emotion. *Speech Commun.*, 40(1-2):161–187, 2003.

[19] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George. Concatenation-Based MIDI-to-Singing Voice Synthesis. In *103rd Meeting of the AES*. New York, 1997.

[20] F. Malfrère, Thierry Dutoit, and Piet Mertens. Automatic prosody generation using suprasegmental unit selection. In *SSW3*, pages 323–328, 1998.

[21] Yoram Meron. *High Quality Singing Synthesis Using the Selection-based Synthesis Scheme*. PhD thesis, University of Tokyo, 1999.

[22] J.L. Miller. *Mandatory processing in speech perception: A case study*, volume Modularity in knowledge representation and natural-language understanding, pages 310–322. J.L. Garfield, www.pallier.org/papers/thesis/thesehtml/node8.html, 1987.

[23] G. Peeters. A large set of audio features for sound description (similarity and classication) in the cuidado project. Technical report, IRCAM, 2004.

[24] G. Peeters and S. McAdams. Instrument sound description in the context of mpeg-7. In *ICMC*, Berlin, Germany, 2000.

[25] C. Pereira and C. Watson. Some acoustic characteristics of emotion. In *Fifth International Conference on Spoken Language Processing, Sydney*, 1998.

[26] H. Pfitzinger. Unsupervised speech morphing between utterances of any speakers. In *AICSST*, pages 545–550, 2004.

[27] Romain Prudon and Christophe d'Alessandro. A selection/concatenation TTS synthesis system: Databases developement, system design, comparative evaluation. In *4th Speech Synthesis Workshop*, Pitlochry, Scotland, 2001.

[28] Prudon R., d'Alessandro C., and Boula de Mareüil. Prosody Synthesis by Unit Selection and Transplantation on Diphones. Santa Monica, USA, 2002.

[29] A. Robel and X. Rodet. Real time signal transposition with envelope preservation in the phase vocoder. In *ICMC*, 2005.

[30] Xavier Rodet. Synthesis and processing of the singing voice. In *Proceedings of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA)*, Leuven, Belgium, November 2002.

[31] J.-L. Rouas, J. Farinas, and F. Pellegrino. Evaluation automatique du débit de parole sur des données multilingues spontannées. 1998.

[32] K.R. Scherer. *Vocal correlates of emotion*, pages 165–197. 1989.

[33] M. Schröder. Emotional speech synthesis–a review. In *Eurospeech, Aalborg*, pages 561–564, DFKI, Saarbrücken, Germany: Institute of Phonetics, University of the Sarland, 2001. schroed@dfki.de.

[34] Diemo Schwarz. A System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 97–102, Verona, Italy, December 2000.

[35] Diemo Schwarz. The CATERPILLAR System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, pages 135–140, London, UK, September 2003.

[36] Diemo Schwarz. *Data-Driven Concatenative Sound Synthesis*. Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris, 2004.

[37] Diemo Schwarz and Matthew Wright. Extensions and Applications of the SDIF Sound Description Interchange Format. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 481–484, Berlin, Germany, August 2000.

[38] Ferréol Soulez, Xavier Rodet, and Diemo Schwarz. Improving Polyphonic and Poly-Instrumental Music to Score Alignment. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, Maryland, USA, October 2003.

[39] Y. Stylianou and O. Cappé. Statistical methods for voice quality transformation. In *EUROSPEECH*, pages 447–450, 1995.

[40] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *ICASSP*, pages 97–100, 2004.

[41] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S.S. Narayanan. Constructing emotional speech synthesizers whith limited speech database. In *ICSLP*, Department of Computer Science and Engineering, Nagoya Istitute of Technology, 2004.

[42] L. Wheeldon. *Do speakers have access to a mental syllabary*, volume Cognition, chapter 50, pages 239–259. www.pallier.org/papers/thesis/thesehtml/node8.html, 1994.

[43] H. Ye and S. Young. High quality voice morphing. In *ICASSP*, 2004.

[44] B. Zellner. Caractérisation du débit de parole en français. In *Journées d'Etude sur la Parole*, Brigitte.zellner@imm.unil.ch, 1998. Faculté de Lettres, Université de Lausanne.