# Première partie

# modèle statistique - projet VIVOS

# Table des matières

**Résumé**

Dans cette présentation, nous décrivons un système de transformation de l'expressivité de la parole. Il vise à modifier l'expressivité d'une phrase neutre, parlée ou synthétisée. La transcription phonétique, le niveau accentuel et les autres informations sur le texte correspondant fournissent une séquence de contextes. Chaque contexte correspond à un jeu de paramètre de transformation acoustique. Ces paramètres changent le long de la phrase et sont utilisés par un vocodeur de phase pour transformer le signal de parole. La relation entre les paramètres de transformation et les contextes est initialisée par un jeu de règles. Un réseau Bayesian transforme progressivement ce modèle à base de règle dans un modèle conduit par les données dans une phase d'apprentissage impliquant une base de données de parole expressive. Le système fonctionne pour le français et pour quelques expressivités. Il est employé à des fins artistiques pour le multimédia, le théâtre et le cinéma.

## 0.1 Introduction

La capacité d'exprimer et d'identifier des émotions par la modulation de caractéristiques de la voix est fondamentale dans la communication humaine. La nature de ces variations contrôlées ou non [Scherer, 1989] provient de plus d'une catégorie (émotions, intentions, attitudes...). Pour désigner cet ensemble, nous utilisons le terme « expressivité » tout en sachant qu'il faut bien distinguer ces catégories. Les systèmes de synthèse de la parole à partir du texte produisent aujourd'hui une parole assez naturelle et intelligible. Dans le domaine artistique, de nombreux compositeurs, metteurs en scène et réalisateurs 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 [Beller, 2007].

Les modèles statistiques ont été utilisés pour la conversion de voix [Hsia *et al.*, 2007], ainsi que pour la synthèse de parole [Bulut *et al.*, 2007, Yamagishi *et al.*, 2005]. Notre système doit pouvoir changer l'expressivité d'une phrase comme le ferait un acteur. C'est pourquoi nous avons construit une base d'exemple en enregistrant des acteurs français. Les enregistrements sont ensuite analysés selon les cinq dimensions de la prosodie [Pfitzinger, 2006] :

- l'intonation (hauteur/mélodie de la voix)
- l'intensité (reliée à l'énergie de la voix)
- le débit de parole (relié à la structure syllabique en français)
- le degré d'articulation (relié aux articulateurs)
- la qualité vocale (reliée à la glotte, absente de cette étude)

Une première étude sur le débit de parole [Beller *et al.*, 2006] a montré l'importance de la connaissance du niveau accentuel des syllabes. Par exemple, les syllabes accentuées durent deux fois plus longtemps que les non-accentuées dans le cas de la joie, alors que toutes les syllabes possèdent à peu près la même durée dans le cas de la peur. Cette information sur les syllabes aident à l'analyse et à la modification de l'expressivité. Un autre exemple de la nécessité de prendre en compte le contexte concerne l'analyse du degré d'articulation. Celui-ci nécessite le tracé d'un triangle vocalique et donc de l'information "voyelle" issue de l'étiquetage phonétique. Cette étape permet notamment de catégoriser les sons de la parole en classes phonétiques dans lesquelles ils vont pouvoir être comparés spectralement. Ainsi le degré d'articulation peut-être mesuré pour toutes les phrases indépendamment de ce qui a été dit et ainsi servir comme descripteur de l'expressivité.

En effet, une difficulté majeure de l'analyse des traits para-linguistiques résident dans l'influence du contenu verbal sur la prononciation. Paradoxalement, l'étude de phénomènes para-verbaux demande une analyse du contenu verbal au préalable, afin de séparer efficacement les effets segmentaux des effets supra-segmentaux. La catégorisation

dépendante du contexte est un outil précieux pour l'analyse des aspects para-linguistiques de la parole. Les unités linguistiques qui composent les enregistrements sont classées suivant leurs contextes phonétiques, leurs niveaux accentuels et d'autres informations symboliques. Des modèles statistiques des paramètres acoustiques de la prosodie sont alors estimés pour chacune des classes contextuelles. Enfin, les modèles engendrés sont comparés entre les classes et leur différences sont reliées aux divers expressivités.

Notre première approche dans la transformation de l'expressivité a été basée sur des règles comme de nombreuses autres laboratoires [Schröder, 2001]. Les variation prosodiques représentées par des facteurs de transposition, de dilatation/expansion temporelle et de gain on été réglées à la main. "Afin de transformer une phrase neutre en phrase joyeuse, transposer les parties voisées d'une octave vers le haut" est un exemple de règles qui ont été écrites à la main et appliquées. De manière à profiter de cette expertise tout en rendant le modèle plus complexe, nous avons décider d'enrichir ces règles grâce à des algorithmes de d'apprentissage automatique. Ainsi, les paramètres de transformation sont appris sur des exemples réels par un réseau bayésien. Un modèle initial à base de règles est en partie modifié en un modèle guidé par les données, selon le nombre d'exemples observés disponibles.

Après une présentation succinte du système, de la base données et des descripteurs utilisés, cette présentation explique comment les paramètres de transformation dépendante du contexte sont inférés par un réseau bayésien et utilisés pour modifier l'expressivité d'une phrase. La phrase neutre à transformer peut provenir directement d'un enregistrement ou bien d'un synthétiseur à partir du texte qui fournira en plus la segmentation phonétique.

## 0.2 Modèle dépendant du contexte

Notre système prend en compte deux niveaux d'information de la parole. D'un côté, la partie linguistique du message, i.e. le texte ainsi que des informations supplémentaires comme l'expressivité, constituent un ensemble de variables discrète (catégorielle) notées dans la suite $S_{variable}$ (représentées par des cercles sur la figure 1). De l'autre côté, la réalisation acoustique du texte, i.e. la parole produite, fournit des données acoustiques continues notées $A_{variable}$ (représentées par des rectangles sur la figure 1). L'étape de segmentation phonétique est fondamentale (et donc vérifiée manuellement) car elle connecte les unités sonores aux descriptions symboliques du texte. Les variables impliquées dans la construction du modèle génératif sont donc précédées d'un $S$ ou d'un $A$ selon qu'elles soient de nature symbolique (discrète) ou acoustique (continue).

### 0.2.1 But d'un modèle génératif

Notre système est conçus pour transformer l'expressivité d'une phrase neutre en une expressivité $S_{exp} = E$ avec un certain degré d'intensité expressive $S_{degree} = D$. Tout d'abord, les informations contextuelles sont extraites pour chaque phone composant la phrase neutre (voir partie 0.4). Cela fournit une séquence temporelle de contextes $C$ (voir partie 0.2.2). Puis deux ensembles de descripteurs acoustiques sont prédits, l'un dans le cas neutre (source) et l'autre dans le cas $E$ (cible). Ces ensembles sont ensuite comparés afin de fournir des facteurs de transformation. Ainsi, le problème revient à inférer des valeurs acoustiques $A$ correspondant à un contexte donné : $S = C_i$, i.e. d'évaluer $P(A|S = C_i)$.

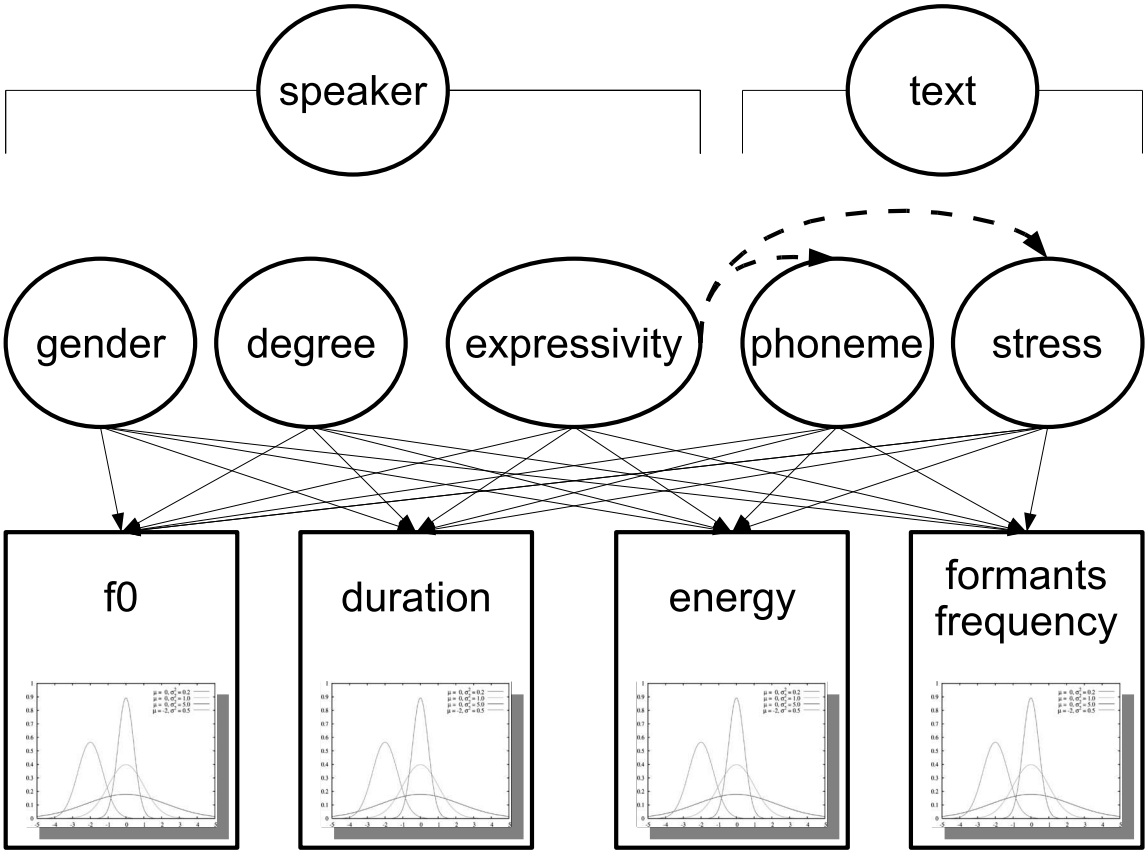


FIG. 1: Réseau bayésien : variables discrètes (cercles) et continues (rectangles), et leurs dépendances.

### 0.2.2 Définition du contexte

Le contexte est défini comme l'ensemble des variables symboliques qui peuvent prendre différents états de vocabulaires fermés. $C_i$ est un exemple de contexte :

$$C_1 = \begin{cases} S_{sexe} & = \text{“}male\text{”} \\ S_{exp} & = \text{“}neutre\text{”} \\ S_{degree} & = \text{“}3\text{”} \\ S_{accent} & = \text{“}non-accentue\text{”} \\ S_{phonem} & = \text{“}/œ/\text{”} \end{cases}$$

Ces variables peuvent découler d'autres variables situées en amont comment $S_{speaker}$ et $S_{text}$.

Le tableau 1 montrent le nombre d'états (*cardinalité*) que peuvent prendre les variables symboliques :

| variable | cardinalité | description |
|---|---|---|
| $S_{sexe}$ | 2 | féminin ou masculin |
| $S_{exp}$ | 15 | *expressivités* (voir partie 0.4.2) |
| $S_{degree}$ | 6 | degré d'intensité *expressive* |
| $S_{accent}$ | 3 | non-accentué, secondairement accentué, accentué |
| $S_{phonem}$ | 34 | phonèmes (code XSAMPA) |

TAB. 1: Noms, cardinalités et descriptions des variables symboliques

Ces variables symboliques sont supposées indépendantes puisque les niveaux phonétique, accentuel, de l'expressivité et de l'identité du locuteur peuvent se réaliser dans n'importe quelle combinaison (ceci est discuté dans la partie 0.4.3). Ainsi l' *Univers* $\mathcal{U}$ de notre modèle dépendant du contexte est composé de 9180 contextes possibles par sexe.

## 0.3 D'un modèle basé sur des règles à un modèle guidé par les données

Pour des utilisations artistiques, le modèle nécessite une certaine flexibilité et doit permettre à l'utilisateur des réglages simples et compréhensibles. Le modèle à base de règles est un bon départ dans ce sens puique ses paramètres sont directement créés par un utilisateur. La contrepartie est une trop grande simplicité du modèle vis à vis de la cardinalité de $\mathcal{U}$. Contrairement à un modèle à base de règles, un modèle guidé par les données peut s'avérer très complexe et, donc, incontrôlable. Il peut aussi être trop proche des données et souffrir d'une faible capacité de généralisation. Un modèle

statistique paramétrique est utilisé car il concilie les avantages des deux méthodes [Yamagishi *et al.*, 2005]. Dans le but de complexifier notre modèle à base de règles, nous employons le paradigme de Bayes. Après une phase d'apprentissage, notre modèle initial est partiellement changé en un modèle statistique guidé par les données, selon le nombre d'observation par contexte. Nous présentons brièvement la transition entre les modèles en respectant l'ordre chronologique de notre approche. (voir [Naïm *et al.*, 2004] pour plus de détails).

### 0.3.1 Modèle à base de règles

Notre première tentative dans la transformation de l'expressivité a été basée sur un ensemble de règles cumulatives créées manuellement selon des exemples enregistrés. Par exemple, chaque "/œ/" *neutre* (contexte $C_1$), est allongé temporellement d'un facteur 1,5 pour être transformer en *tristesse extravertie* (contexte $C_2$), puis à nouveau par un facteur 1,8 s'il est accentué. Une telle règle peut s'écrire :

$$\begin{aligned} A_{duree}(C_3) &= 1.8 \times A_{duree}(C_2) \\ &= 1.8 \times 1.5 \times A_{duree}(C_1) \end{aligned}$$

Mais la cardinalité de l'Univers $\mathcal{U}$ rend la tâche de construction du modèle très complexe. C'est la raison pour laquelle nous avons choisi un paradigme d'apprentissage sur un corpus d'exemples.

### 0.3.2 Approche fréquentiste

La base de données employées (corpus d'exemples) est décrite dans la partie 0.4. soient :

- $X = \{X_{(l)}\}_{l=1..N}$ l'ensemble de $N$ données observées
- $\theta$ les paramètres du modèle
- $S_{acous}$ une variable acoustique discrète (plutôt que continue) pour l'explication.

Si toutes les variables sont complètement observées, c'est à dire que nous possédons une mesure des variables acoustiques $A$ dans chaque contexte $S \in \mathcal{U}$, une méthode intuitive est de mesurer la probabilité d'un événement $(S_{acous} = S_j)$ par la fréquence d'apparition de cet événement dans un contexte $(S = C_i)$. Cette approche, appelé maximum de vraisemblance (MV), donne :

$$\hat{P}(S_{acous} = S_j | S = C_i) = \hat{\theta}^{MV}_{i,j} = \frac{N_{i,j}}{\sum_j N_{i,j}} \tag{1}$$

où $N_{i,j}$ est le nombre de fois que $S_{acous} = S_j$ dans le contexte $S = C_i$.

### 0.3.3 Approche bayésienne

Ce formalisme nous permet de résonner sur des probabilités selon des conditions de certitude. Similaire à l'approche fréquentiste, il prend en compte un critère objectif supplémentaire d'optimisation qui incorpore une distribution à priori sur la quantité que l'on souhaite estimer. Cela consiste à chercher les paramètres $\theta$ du modèle les plus probables, sachant que les données on été observées et partant d'un à priori sur ces paramètres. Cette approche, appelée espérance à posteriori (EAP), donne :

$$\hat{P}(S_{acous} = S_j | S = C_i) = \hat{\theta}_{i,j}^{EAP} = \frac{N_{i,j} + \alpha_{i,j}}{\sum_j (N_{i,j} + \alpha_{i,j})} \tag{2}$$

où $\alpha_k$ sont les paramètres d'une distribution de Dirichlet associée à l'à priori $P(S_{acous} = S_j | S = C_i)$. $\alpha_k$ sont les paramètres qui contrôlent les poids attribués aux règles et aux données dans le modèle final. Si $\alpha_{i,j} \to \infty$, le modèle final sera complètement influencé par les règles données à priori et si $\alpha_{i,j} \to 0$, le modèle final sera complètement influencé par les données. Ainsi, le poids des règles dans le modèle final est définit par un rapport entre le nombre de cas simulés à priori (contrôlable) et le nombre de cas observés dans les données (fixé). Il existe aussi l'approche du maximum à posteriri (MAP).

### 0.3.4 Réseau bayésien

Les réseaux bayésiens ont été utilisés dans différentes domaines du traitement de la parole. Des classifieurs bayésiens naïfs et des réseaux bayésiens dynamiques sont utilisés en reconnaissance des émotions [Ball, 2003]. Un réseau bayésien modèle des dépendances entre des variables discrètes et continues. il se compose d'une description qualitative représentée par un graphe et d'une description quantitative représentée par une fonction de densité de probabilité généralisée (ou jointe) GPDF :

$$GPDF = P(A, S) \tag{3}$$

#### Partie qualitative : Modèle graphique

La structure du modèle graphique est ici donnée arbitrairement (elle peut être apprise) et présentée dans la figure 1. Cette vision qualitative du modèle statistique montrent les variables impliquées durant les phases d'apprentissage et d'inférence. Les cercles représentent les variables discrètes composant le contexte symbolique. Les rectangles représentent les variables continues qui sont des vecteurs de valeurs caractéristiques calculées sur les descripteurs acoustiques dynamiques (voir partie 0.5.1). Les flèches représentent les dépendances entre les variables.

**Partie quantitative : fonction de densité de probabilité généralisée** $P(A, S)$

La fonction de densité de probabilité généralisée $GPDF$ quantifie toutes les dépendances entre les variables du réseau bayésien. Chaque variable continue se voit attribuée une distribution gaussienne linéairement conditionnelle (LCG), dépendante de la configuration de ses variables parentes discrètes. La $GPDF$ est estimée grâce à la loi de Bayes :

$$P(A, S) = P(A|S)P(S) \tag{4}$$

Une fois la $GPDF$ estimée (phase d'apprentissage), les distributions LCG $P(A|S = C_i)$ des variables acoustiques sont inférées en utilisant l'inverse de l'équation 4 (phase d'inférence).

## 0.4 Base de données

De manière à estimer la $GPDF$, nous avons enregistré une base de données expressives en français. Elle présente 3996 contextes par sexe. Donc $\mathcal{U}_{observed}$ n'est pas exhaustive et ne couvre que moins de la moitié de l'Univers $\mathcal{U}$. Ce manque de données soulève des difficultés lorsqu'une nouvelle phrase(contexte) est présentée au système. Ce problème est considéré et partiellement résolu dans la partie 0.5.3.

### 0.4.1 Enregistrements

La base de données est composée des enregistrements de quatre acteurs ($S_{speaker}$), deux hommes et deux femmes ($S_{sexe}$), durant chacun approximativement une heure et demi. Ces acteurs ont tous été enregistrés dans les mêmes conditions professionnelles, ont tous suivi la même procédure et nt tous lu le même texte : Dix phrases ($S_{text}$) extraites d'un corpus de textes phonétiquement équilibrés [Combescure, 1981] et marquées prosodiquement grâce à la ponctuation et au soulignage de syllabes accentuées.

### 0.4.2 Expressivités

Les expressivités choisies ($S_{exp}$) sont des émotions actées : *neutre, colères introvertie et extravertie, joies introvertie et extravertie, peurs introvertie et extravertie, tristesses introvertie et extravertie*, et *les surprises positive et négative, le dégoût, la discrétion, l'excitation et la confusion*. Chaque phrase a été prononcée dans toutes les expressivités. De plus, dans le cas des émotions actées, chaque phrase a été répétée six fois avec une degré d'intensité expressive ($S_{degree}$) croissant. Finalement, le corpus se compose d'approximativement 550 phrases par acteur. Des *fillers* ont aussi été enregistré pour chaque expressivité.

### 0.4.3 Segmentation phonétique

Le premier pas de l'analyse de ces données est la segmentation phonétique des phrases en phones ($S_{phonem}$). La méthode de segmentation automatique employée [Morris, 2006] est classique et repose sur des chaînes de Markov cachées entraînées sur une base de données de parole neutre multi-locuteur [Lamel *et al.*, ]. Cette segmentation initiale a ensuite été corrigée manuellement par des phonéticiens avec l'outil wavesurfer [Sjölander et Beskow, 2000]. La réalisation phonétique, dans le cas expressif, peut être assez lointaine de la chaîne phonémique automatiquement prédite à partir du texte. En effet, la réalisation phonétique d'une phrase expressive est influencée par l'expressivité comme le montre la partie 0.6.1.

### 0.4.4 Segmentation prosodique

La segmentation phonétique fournit des labels XSAMPA utilisés par des traitements ultérieurs qui on pout ut de définir les frontières prosodiques et les durées d'autres types d'unités sonores : *syllabe, groupe prosodique, groupe de souffle* et *phrase*. Un syllabifieur (syllabificateur ?) par règles définit les syllables à partir de la de la chaîne phonétique. Les frontières prosodiques sont alors automatiquement définies par des prédictions du niveau d'accent des syllabes à partir du texte, corrigées manuellement ($S_{accent}$). Ces descriptions et d'autres descripteurs symboliques définissant les places relatives des unités les unes par rapport aux autres sont stockés dans des fichiers XML qui permettent l'archivage des relations hiérarchiques.

### 0.4.5 Descripteurs acoustiques

Chaque unité étiquetée est un segment de parole analysé. Les descripteurs dynamiques sont des quantités d'analyses acoustiques évoluant durant l'horizon temporel de chaque unité.

#### fréquence fondamentale et énergie

La courbe de hauteur, la mélodie ou encore le contour intonatif est un indice perceptuel fondamental dans l'expressivité. La fréquence fondamentale ($A_{f0}$) est calculée par l'algorithme YIN [de Cheveigné et Kawahara, 2002]. Cet algorithme calcule aussi l'énergie ($A_{energy}$) et un indice de voisement pour chaque trame de signal. La fréquence fondamentale est interpolée entre les segments voisés dont les frontières sont déterminées par seuillage de l'indice de voisement (voir figure 2).

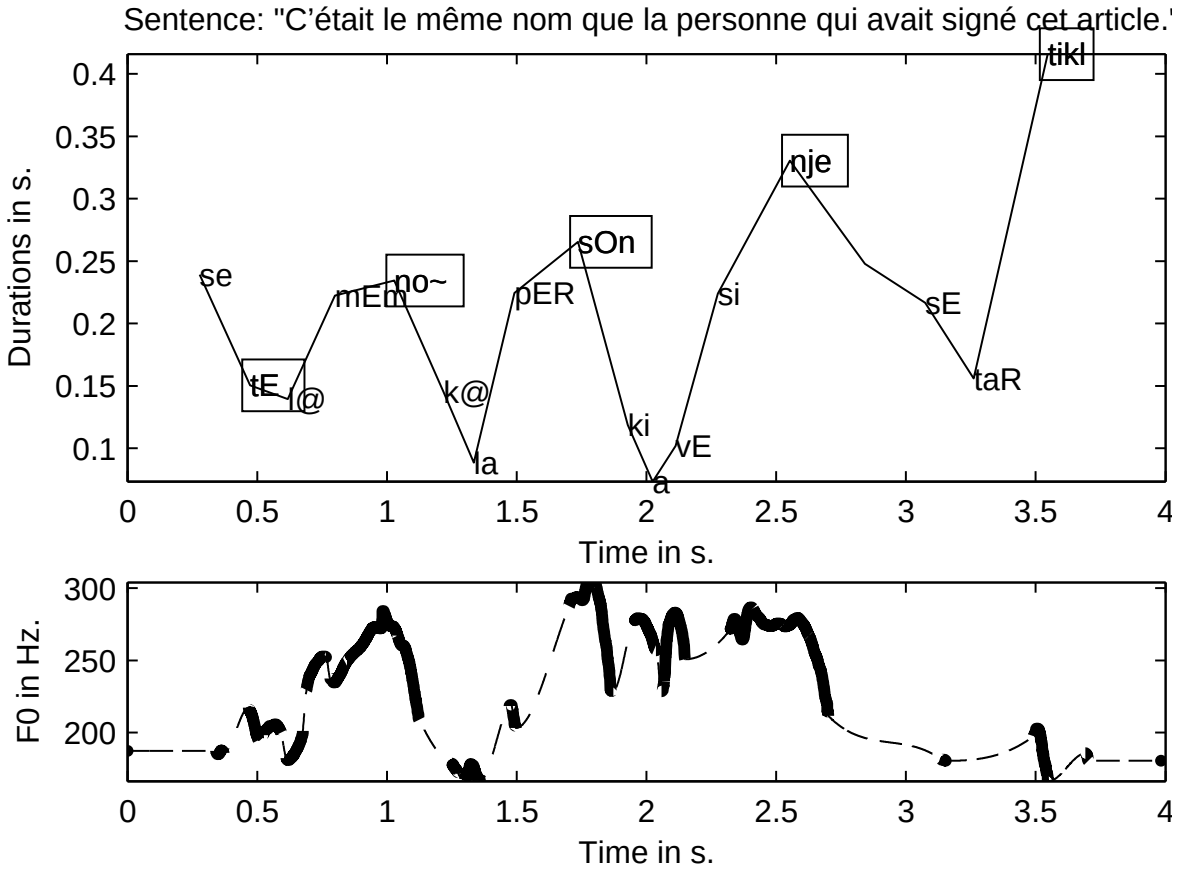


FIG. 2: Durées et fréquence fondamentale des syllabes d'une phrase prononcée par un homme simulant la joie extravertie : les syllabes encadrées sont accentuées.

#### Débit de parole

Le débit de parole local est défini à partir des durées des syllabes [Beller *et al.*, 2006] (voir figure 2). Contrairement à la définition consensuelle du débit de parole moyen calculé sur la phrase entière et dont l'unité est le nombre de syllabes par seconde, nous gardons la durée des syllabes en tant qu'unité. Les syllabes accentuées possèdent souvent une durée plus longue que les syllabes non-accentuées . Ainsi la courbe de débit locale définie directement par ces durées présente des maxima situés sur les accents. Cela permet une visualisation de l'évolution du débit de parole au cours de la phrase. Une décélération correspond à une montée de la courbe tandis qu'une accélération correspond à une chute de la courbe.

#### Fréquence des formants

Les paramètres des formants sont calculés par un algorithme d'estimation des trajectoires de formants [Beller, 2007]. En premier, la méthode trouve les pôles d'un modèle auto-régressif, estimé par LPC du signal découpé en trames temporelles fines et fenêtrées. Puis elle définit les pôles les plus importants grâce à une combinaison de critères sur les paramètres dont le délai de groupe [Murthy *et al.*, 1989]. Enfin, elle fait correspondre certains de ces pôles aux formants tout en assurant que les trajectoires des formants soient continues dans le plan temps-fréquence. Les trajectoires sont décodées récursivement grâce à la programmation dynamique.

## 0.5 Transformation

Une fois la phase d'apprentissage terminée (la $GPDF$ estimée), une nouvelle phrase peut être présentée. Une séquence de contextes symboliques est défini à partir du texte, de l'accentuation et de l'expressivité désirée. Les variables acoustiques sont alors inférées pour chaque phonème selon son contexte. La démarche utilisée est résumée dans l'algorithme 1.

### 0.5.1 Modèle temporel

L'évolution de chaque descripteur dynamique est modélisée temporellement sur l'horizon temporel de chaque unité. Les paramètres de ce modèle se présentent sous la forme d'un vecteur de valeurs caractéristiques : (voir figure 3).

- valeurs initiales, finales, au milieu, minimum, maximum et écart absolu.
- moyennes arithmetic, geometric et écart type
- Centres de gravité et d'anti-gravité temporels donnant les lieux de la plus importante élévation et dépression de la courbe.

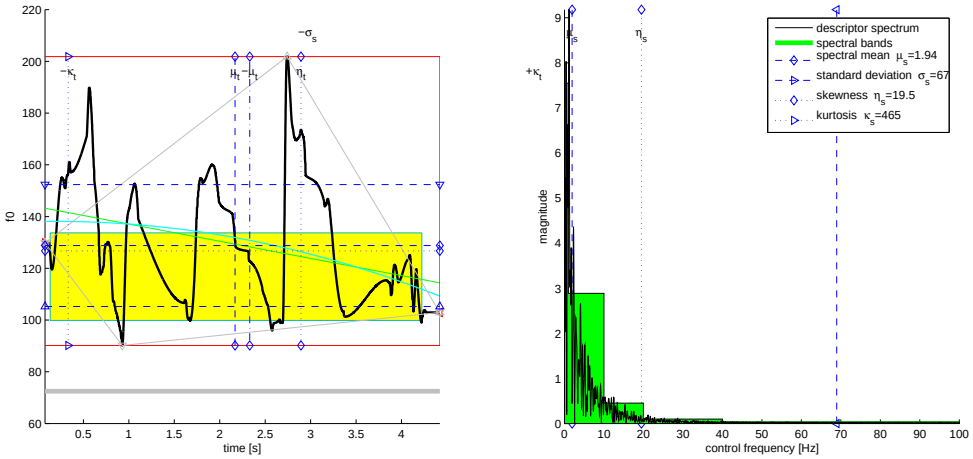


FIG. 3: Exemple d'un vecteur de valeurs caractéristiques modélisant l'évolution temporelle de la fréquence fondamentale (en [Hz]) sur la durée d'une voyelle (en [s]) prononcée avec colère introvertie.

- approximation polynômial de Legendre du $2^{nd}$ ordre donnant la pente et la courbure
- Point d'inflexion de la courbe correspondant à la valeur cible, prise à l'instant ou la dérivée du polynôme du $2^{nd}$ ordre s'annule ou au milieu si l'évolution est linéaire.
- Spectre de Fourier et centroïde spectral du descripteur, Le *Jitter* est ici mesuré par la déviation du centre de gravité du spectre de Fourier de f0. Le *Shimmer* est mesuré similairement sur l'énergie.

Certaines de ces valeurs caractéristiques ne sont utilisées que pour l'analyse. Pour le moment, seule les valeurs aux points d'inflexion sont prises en compte par le réseau bayésien.

### 0.5.2 Inference

Deux inférences fournissent deux possibles réalisations acoustiques d'une même phrase prononcée de manière neutre ou prononcée de manière expressive. La comparaison des ces deux ensembles de valeurs acoustiques fournit des facteurs de transposition, de dilatation/compression temporelle, de gain et de réassignement spectral qui évoluent durant la phrase puisque le contexte change à chaque phone. Après une phase de lissage de ces paramètres de transformation, un vocodeur de phase [Bogaards *et al.*, 2004] transforme le signal de la phrase neutre selon ces paramètres dynamiques.

### 0.5.3 Contextes non observés

Une nouvelle phrase peut présenter un contexte qui n'a pas été observé durant la phase d'apprentissage puisque nos données ne couvre pas tout l'Univers $\mathcal{U}$. Dans ce cas, le modèle génératif doit tout de même fournir des paramètres de transformation. Deux phases d'inférence sont alors nécessaires. La première utilise directement les valeurs des variables acoustiques mesurées sur la phrase neutre (à transformer) afin d'inférer la séquence de contextes la plus probable. L'expressivité et le degré d'intensité désirés sont alors ajoutés/modifiés à ces contextes. La seconde phase d'inférence permet alors de prédire des données acoustiques. La phase d'apprentissage ne possède pas pour l'instant la capacité de généralisation à tous les contextes. C'est une question cruciale qui dépasse le propose de cette présentation mais qu'il est important de traiter. Pour le moment, la solution présentée ici permet de déduire des paramètres de transformation pour tous

les contextes possibles en procédant par analogie.

**→Initialisation**

- $GPDF$ estimée (phase d'apprentissage terminée) ;
- contextes observés : $\mathcal{U}_{observed}$ ;
- nouvelle phrase neutre [N] (audio et texte) ;
- expressivité [E] et degré d'intensité expressive [D] désirés;

**→Analyses**

- segmentation en $P$ phones ;
- définition d'une séquence de contextes $\{C_N(t)\}_{t\in[1:P]}$ ;
- analyses acoustiques ;
- calcul des valeurs caractéristiques ;

**→Inference des variables acoustiques**

**for** $t \in [1 : P]$ **do**

- vérifier si le contexte a été observé ou pas :

**if** $C_N(t) \in \mathcal{U}_{observed}$ **then**

inference des variables acoustiques neutres $A_N(t)$ correspondant au contexte $C_N(t)$ ;

**else**

inference des contextes $C_N(t)$ correspondant aux données acoustiques $A_N(t)$ ;

**end**

- Ajout de l'expressivité désirée au contexte $C_E(t)$ :

$$C_E(t) = C_N(t)$$

$$C_E(t) = \begin{cases} S_{exp} & = \text{E} \\ S_{degree} & = \text{D} \end{cases}$$

- inference des variables acoustiques expressives $A_E(t)$ correspondant au contexte $C_E(t)$ ;

**end**

**→Paramètres de transformation**

- Calcul des paramètres de transformation $T_{N\to E}$ de $A_E$ et $A_N$
- Lissage des paramètres de transformation $T_{N\to E}$;

**→Transformation du signal de parole**

- transposition dynamique ;
- dilation/compression temporelle dynamique ;
- gain dynamique ;
- réassignement spectral dynamique ;

**Algorithm 1**: Algorithme de transformation d'une nouvelle phrase

Cette procédure mènent à une prédiction de paramètres de transformation du signal pour n'importe quel contexte symbolique observé dans la phase d'apprentissage ou non, par un principe d'analogie [van Santen *et al.*, 2003].

## 0.6 Discussions

L'apprentissage bayésien possède de nombreux avantages par rapport à un système à base de règles. Notre précédent modèle à base de règles est toujours utilisé pour l'initialisation du modèle final. L'application de la loi de Bayes permet le calcul de distributions des paramètres acoustiques qui concordent mieux aux données selon le nombre d'observations par contexte. La boîte à outil Matlab® pour les réseaux bayésiens [Murphy, 2001] calcule efficacement les distributions de probabilités conditionnelles des variables discrètes et continues. L'hétérogénéité de la nature de ces variables confère au modèle la capacité d'être dépendant du contexte. Cette approche bayésienne soulève de nombreuses questions concernant l'indépendance des variables entre elles, comme nous l'avons souligné dans la partie 0.4.3.

### 0.6.1 Interdépendance des variables dans le contexte

Les phonéticiens qui ont corrigés la segmentation phonétique de la base de données, ont observés que pour certaines expressivités, des phonèmes étaient absents, ajoutés ou différents que ceux fournis par la segmentation automatique (un "/E/" ouvert peut sonner comme un "/œ/", par exemple). Même si le même texte a été prononcé, les fréquences d'apparition des phones $S_{phonem}$ pour chaque expressivité sont différentes. $P(S_{phonem}|S_{exp})$ a été estimée en ajoutant une dépendance (flèche en pointillé sur la figure 1) dans le graphe : $S_{exp} \rightarrow S_{phonem}$. Les fréquences d'apparition par phrase de quelques classes phonologiques ont été mesurées sur les corpus d'un acteur et d'une actrice réunis et sont représentées par la hauteur des barres de la figure 4. Cela montre que l'expressivité influence la prononciation d'un texte. Une étude poussée sur les correspondances entre les chaînes phonémiques désirées et les chaînes phonétiques réalisées fournirait plus d'information.

Une interdépendance contextuelle similaire est observable pour le niveau accentuel : $S_{exp} \rightarrow S_{accent}$ (flèche en pointillé sur la figure 1). Par exemple dans le cas de la colère extravertie, presque chaque syllabe peut être perçue comme accentuée, car elles sont séparées par des césures.

### 0.6.2 Interdépendances entre variables acoustiques

Un second type d'interdépendance important apparaît entre les variables acoustiques. Par example, la variance de la fréquence fondamentale est fortement corrélée au débit de parole. Le degré d'articulation est diminué lorsque la parole est accélérée, bien que le contraire a été observé pour quelques expressivités comme la colère extravertie [Beller *et al.*, 2006]. Ainsi des relations comme $A_{f0} \rightarrow A_{duration}$ doivent être ajoutée au modèle.

### 0.6.3 Dépendance entre deux contextes successifs

Les chaînes de Markov cachées sont largement répandues en parole et partagent le même formalisme que les réseaux bayésiens (modèles graphiques). Il a été montré que la connaissance des probabilités de transition entre les phonèmes augmente les performances des systèmes de reconnaissance automatique. Si bien que nous espérons une augmentation de la qualité de la prédiction en connectant deux contextes successifs : $S(i-1) \rightarrow S(i)$. Ainsi les phénomènes de coarticulation pourront aussi être modélisés par un réseau bayésien dynamique.

### 0.6.4 Variable expressivité : discrète ou continue ?

Finalement, la nature de $S_{exp}$ peut être discutée car il esxiste plusieurs représentations de l'expressivité dont certaines sont catégorielles (discrète) et d'autres sont dimensionnelles (continues) [Schröder, 2001]. Notre choix n'a pas été arrêté et reste une question ouverte. Cependant, nous allons rendre la variable discrète $S_{degree}$, continue. Ainsi les

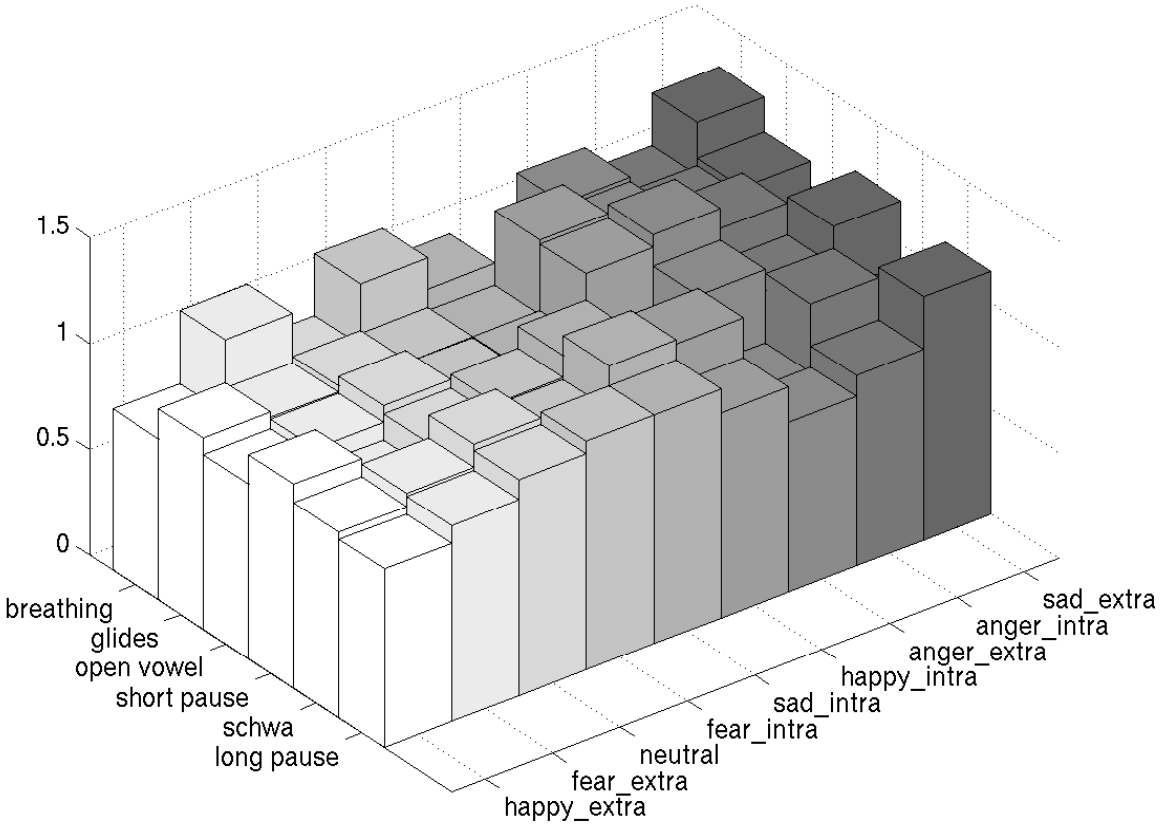


FIG. 4: Fréquences d'apparitions par phrase moyennes de classes phonologiques en fonction de l'expressivité.

distribution LCG des paramètres acoustiques vont devenir linéairement dépendantes ($W_i$) de la valeur du degré d'intensité expressive :

$$P(A_{f0}|S_{degree} = d, S = C_i) = \mathcal{N}(\mu_i + W_i \times d, \sigma_i) \tag{5}$$

## 0.7 Conclusion

Dans cette présentation, nous avons décrit un système de transformation de l'expressivité de la parole. Un modèle génératif statistique est appris sur une base de données de parole expressive multi-locuteur. Les paramètres des transformations acoustiques varient dans le temps et sont dépendants des contextes symboliques extraits du texte et d'une définition de l'état du locuteur. Il a été montré comment un réseau bayésien réalise le passage entre une modèle à base de règles et un modèle guidé par les données. Même si toutes les possibilités et améliorations évoquées dans cette présentation n'ont pas encore été testées, il semble que la solution proposée soit suffisamment transparente, interprétable et adéquate au problème, pour être retenue. Le système fonctionne pour quelques expressivités en français. Des exemples sonores sont disponibles à l'adresse suivante : http://recherche.ircam.fr/equipes/analyse-synthese/beller. Nos travaux futurs comprennent la réalisation des prospectives évoquées précédemment, l'introduction de la qualité vocale dans le modèle et ,enfin, l'évaluation des résultats par des tests d'écoutes.

# Bibliographie


[Ball, 2003] Ball, E. (2003). A bayesian heart : computer recognition and simulation of emotion. *Dans* Trappl R, Petta P, P. S., éditeur : *Emotions in humans and artifacts*. Cambridge, The MIT Press.

[Beller, 2007] Beller, G. (2007). Influence de l'expressivité sur le degré d'articulation. *Dans RJCP, Rencontres Jeunes Chercheurs de la Parole*.

[Beller *et al.*, 2006] Beller, G., Schwarz, D., Hueber, T. et Rodet, X. (2006). Speech rates in french expressive speech. *Dans Speech Prosody*, Dresden. SproSig, ISCA.

[Bogaards *et al.*, 2004] Bogaards, N., Roebel, A. et Rodet, X. (2004). Sound analysis and processing with audiosculpt 2. *Dans International Computer Music Conference (ICMC)*, Miami, USA.

[Bulut *et al.*, 2007] Bulut, M., Lee, S. et Narayanan, S. (2007). a statistical approach for modeling prosody features using pos tags for emotional speech synthesis. *Dans ICASSP*.

[Combescure, 1981] Combescure, P. (1981). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56:34–38.

[de Cheveigné et Kawahara, 2002] de Cheveigné, A. et Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *JASA*, 111:1917–1930.

[Hsia *et al.*, 2007] Hsia, C.-C., Wu, C.-H., et Wu, J.-Q. (2007). conversion function clustering and selection for expressive voice conversion. *Dans ICASSP*.

[Lamel *et al.*, ] Lamel, L., Gauvain, J. et Eskenazi, M. Bref, a large vocabulary spoken corpus for french.

[Morris, 2006] Morris, A. (2006). Automatic segmentation. Rapport technique, IRCAM.

[Murphy, 2001] Murphy, K. (2001). The bayes net toolbox for matlab. *Dans Computing Science and Statistics*, volume 33.

[Murthy *et al.*, 1989] Murthy, H., Murthy, K. M. et Yegnanarayana, B. (1989). Formant extraction from phase using weighted group delay function. *Dans Electronics Letters*, volume 25, pages 1609–1611. IEE.

[Naïm *et al.*, 2004] Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O. et Becker, A. (2004). *Réseaux bayésiens*. Eyrolles, Paris.

[Pfitzinger, 2006] Pfitzinger, H. (2006). Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction. *Dans* Hoffmann, R. ; Mixdorff, H., éditeur : *Speech Prosody*, numéro 40 de Abstract Book, pages 6–9, Dresden.

[Scherer, 1989] Scherer, K. (1989). *Handbook of Psychophysiology : Emotion and Social Behavior*, chapitre Vocal correlates of emotion, pages 165–197. London, Wiley.

[Schröder, 2001] Schröder, M. (2001). Emotional speech synthesis–a review. *Dans Eurospeech, Aalborg*, pages 561–564, DFKI, Saarbrücken, Germany : Institute of Phonetics, University of the Sarland.

[Sjölander et Beskow, 2000] Sjölander, K. et Beskow, J. (2000). Wavesurfer - an open source speech tool. *Dans International Conference on Spoken Language Processing*.

[van Santen *et al.*, 2003] van Santen, J., Black, L., Cohen, G., Kain, A., Klabbers, E., Mishra, T., de Villiers, J. et Niu, X. (2003). Applications of computer generated expressive speech for communication disorders. *Dans EUROSPEECH*.

[Yamagishi *et al.*, 2005] Yamagishi, J., Onishi, K., Masuko, T. et Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. *Dans IEICE Trans. on Inf. & Syst.*, volume E88-D, pages 503–509.