

Caractérisation de l'environnement musical dans les documents audiovisuels

Hélène Lachambre

Directrice de recherche : Régine André-Obrecht
Co-encadrant : Julien Pinquier



8 décembre 2009



Université
de Toulouse

Plan

- 1 Problématique
- 2 Monophonique / Polyphonique
- 3 Détection du chant
- 4 Conclusion

Outline

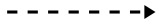
- 1 Problématique
- 2 Monophonique / Polyphonique
- 3 Détection du chant
- 4 Conclusion

Contexte général

- Indexation multimédia
 - Recherche d'informations
 - Navigation
- Analyse de l'audio
- Indexation de grands flux audio par la musique
- Indexation de morceaux de musique

Indexation de la musique

Musique



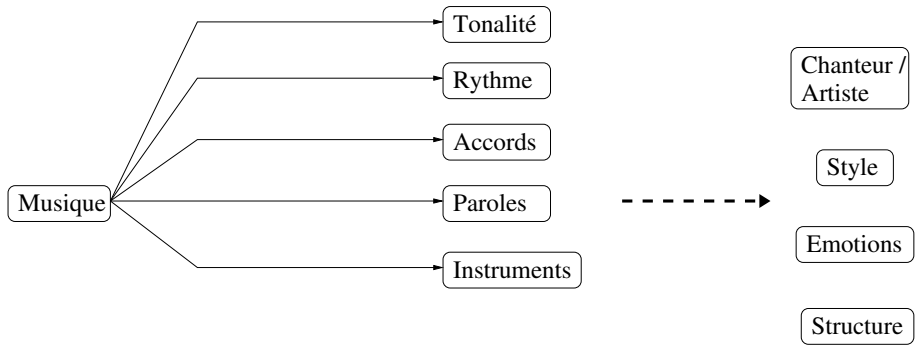
Chanteur /
Artiste

Style

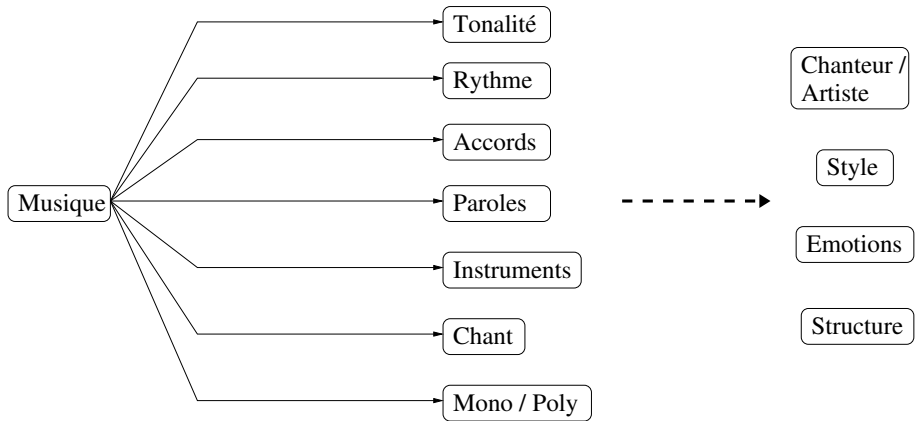
Emotions

Structure

Indexation de la musique



Indexation de la musique



Monophonique / Polyphonique

Problématique

À l'instant t , combien y a-t-il de sources harmoniques ?

Monophonique / Polyphonique

Problématique

À l'instant t , combien y a-t-il de sources harmoniques ?

Une source – Notre définition

Une **source** = Une **note** jouée par un **instrument** (chanteur)

- Une seule source ? → **Monophonie**
 - Flûte, chant
- Plusieurs sources ? → **Polyphonie**
 - Harmonica, piano, orchestre

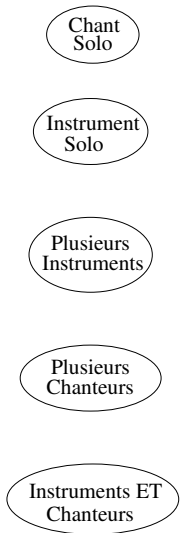
Détection du chant

Problématique

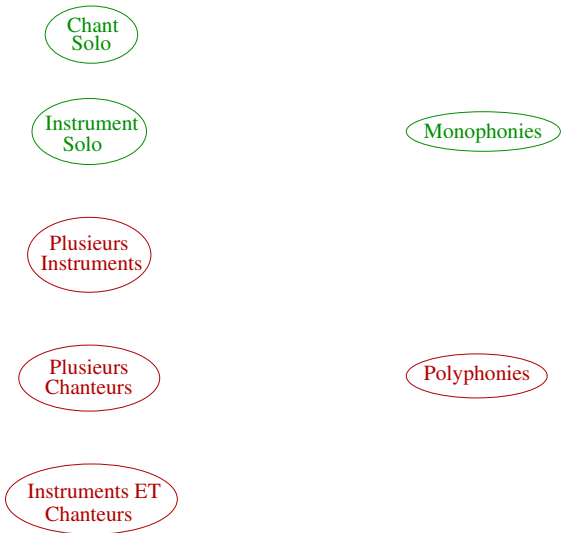
À quels instants y a-t-il du chant ?

Le chant : Parole ? Musique ?

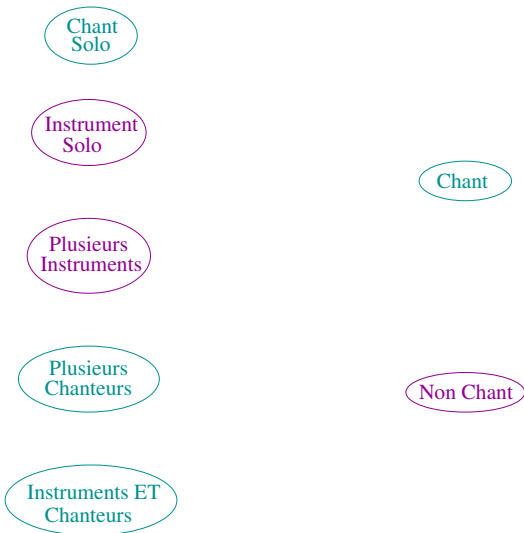
Les classes et sous-classes



Les classes et sous-classes



Les classes et sous-classes



Corpus

- Corpus **équilibré** entre les classes et sous-classes
- Durée d'apprentissage : **25 s / sous-classe** (5*5 extraits)
- Durée totale de test : 18 minutes
- Corpus de test **varié** : Styles (Classique, Rock, Variété, Rap, . . .), Effectifs
- Métrique utilisée : Taux global d'erreur

Outline

- 1 Problématique
- 2 Monophonique / Polyphonique
 - État de l'art
 - Notre approche
 - Paramètres
 - Modélisation - Distributions de Weibull bivariées
 - Expériences et résultats
- 3 Détection du chant
- 4 Conclusion

État de l'art (1/3) [Essid, Richard, Bertrand 2005]

Application : Reconnaissance d'instrument

Contexte : Nombre d'instruments \Rightarrow Nombre minimum de notes

- Paramètres : paramètres temporels, spectraux et MFCC + Sélection de paramètres
- Modélisation : GMM
- Corpus : Jazz
- Performances : 91 % de bonne reconnaissance d'instruments

État de l'art (2/3) [Tsai, Liao, Lai 2008]

Application : Reconnaissance du chanteur

Contexte : Chant solo vs Chant duo

- Paramètres : MFCC
- Modélisation : GMM
- Corpus : Enregistrements « maison »
- Performances : Accuracy = 96 %

État de l'art (3/3) [Smit, Ellis 2007]

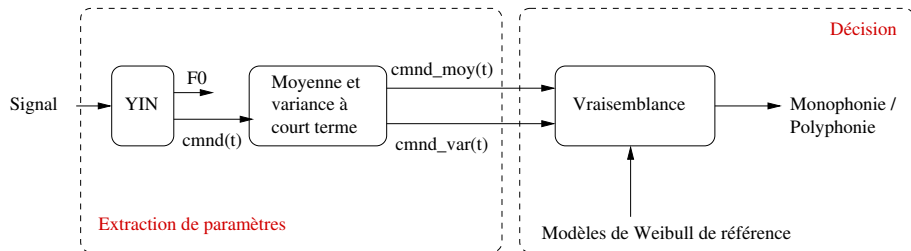
Application : Query by Singing

Contexte : Chant solo vs Polyphonies

- Méthode de référence : MFCC + GMM
- Corpus : Musique Folk et classique
- Performances : Précision = 70 %, Rappel = 50 %

- Méthode proposée : Recherche d'une fonction périodique unique pour modéliser une trame
- Performances : Précision = 70 %, Rappel = 70 %

Système de classification – Approche probabiliste



YIN : estimateur de fréquence fondamentale [de Cheveigné, Kahawara 2002]

La fonction de différence $d_t(\tau)$

$$d_t(\tau) = \sum_{k=1}^N (x_k - x_{k+\tau})^2 \quad (1)$$

Avec : x le signal, τ le décalage et t l'indice de la trame

YIN : estimateur de fréquence fondamentale [de Cheveigné, Kahawara 2002]

La fonction de différence $d_t(\tau)$

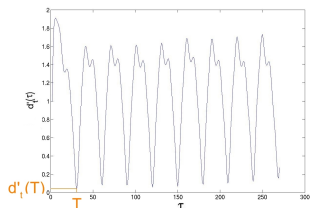
$$d_t(\tau) = \sum_{k=1}^N (x_k - x_{k+\tau})^2 \quad (1)$$

Le « Cumulative Mean Normalised Difference » $d'_t(\tau)$

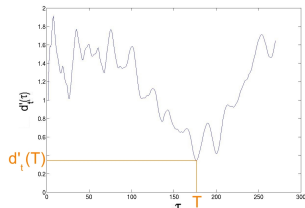
$$d'_t(\tau) = \begin{cases} 1 & \text{si } \tau = 0 \\ d_t(\tau) / \left[\frac{1}{\tau} \cdot \sum_{k=1}^{\tau} d_t(k) \right] & \text{sinon} \end{cases} \quad (2)$$

Avec : x le signal, τ le décalage et t l'indice de la trame

Les paramètres de l'étude



(a) Musique monophonique

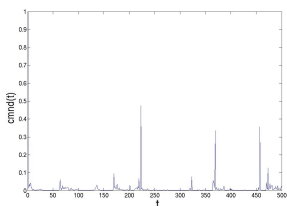


(b) Musique polyphonique

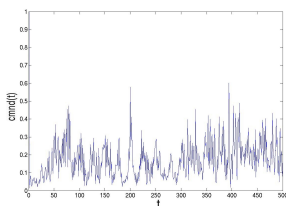
FIG.: $d'_t(\tau)$ pour une trame de signal (20 ms).

- F_0 est donnée par l'*indice* T du minimum de $d'_t(\tau)$
- Posons $cmd(t) = d'_t(T)$
- $cmd(t)$ est un « *indice de confiance* »

Les paramètres de l'étude



(a) Musique monophonique

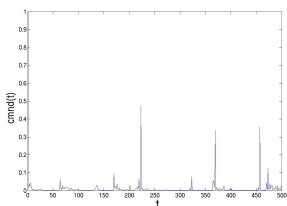


(b) Musique polyphonique

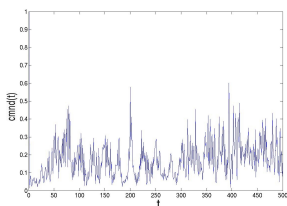
FIG.: Valeurs de $cmnd(t)$ pour 5 secondes de signal.

- Monophonie : $cmnd(t)$ est **bas** et **varie peu**
- Polyphonie : $cmnd(t)$ est **élevé** et **varie plus**

Les paramètres de l'étude



(a) Musique monophonique



(b) Musique polyphonique

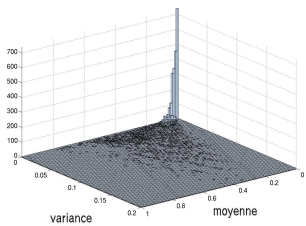
FIG.: Valeurs de $cmnd(t)$ pour 5 secondes de signal.

- Monophonie : $cmnd(t)$ est **bas** et **varie peu**
- Polyphonie : $cmnd(t)$ est **élevé** et **varie plus**

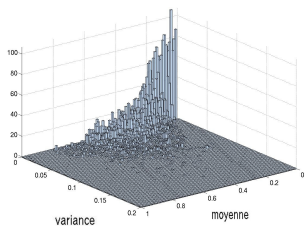
Les paramètres sont :

- La **moyenne court terme** de $cmnd(t)$: $cmnd_{moy}(t)$
- La **variance court terme** de $cmnd(t)$: $cmnd_{var}(t)$

Les paramètres de l'étude



(a) Monophonies

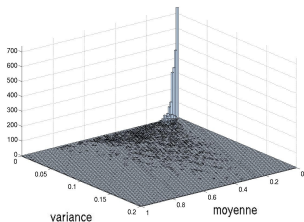


(b) Polyphonies

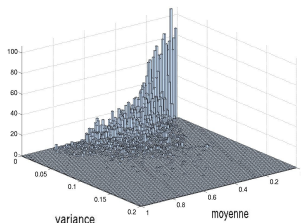
FIG.: Répartition bivarée du couple $(cmnd_{moy}, cmnd_{var})$ - 5000 et 7500 observations.

- $cmnd_{moy}$: moyenne sur 5 trames
- $cmnd_{var}$: variance sur 5 trames

Les paramètres de l'étude



(a) Monophonies



(b) Polyphonies

FIG.: Répartition bivarée du couple $(cmnd_{moy}, cmnd_{var})$ - 5000 et 7500 observations.

- $cmnd_{moy}$: moyenne sur 5 trames
- $cmnd_{var}$: variance sur 5 trames

Modélisation

⇒ Modélisation par des distributions de Weibull bivariées.

La distribution de Weibull (univariée)

Fonction de répartition

$$F(x) = 1 - e^{-(x/\theta)^\beta} \quad (3)$$

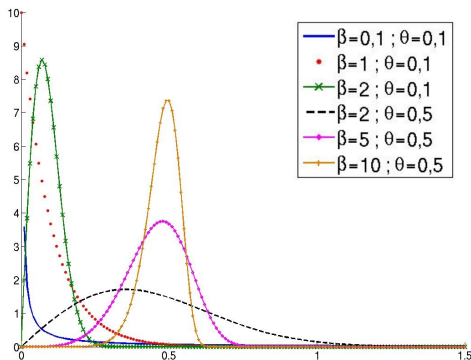


FIG.: Densités de probabilité d'une fonction de Weibull univariée, pour différentes valeurs de paramètres d'échelle θ et de forme β .

La distribution de Weibull bivariée [Hougaard 1986], [Lu, Bhattacharyya 1990]

Fonction de répartition

$$F(x, y) = 1 - \exp \left(- \left[\left(\frac{x}{\theta_1} \right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2} \right)^{\frac{\beta_2}{\delta}} \right]^{\delta} \right) \quad (4)$$

Avec :

- β_1 et β_2 les paramètres de **forme**,
- θ_1 et θ_2 les paramètres d'**échelle**,
- δ le paramètre de **corrélation**.

La distribution de Weibull bivariée [Hougaard 1986], [Lu, Bhattacharyya 1990]

Fonction de répartition

$$F(x, y) = 1 - \exp \left(- \left[\left(\frac{x}{\theta_1} \right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2} \right)^{\frac{\beta_2}{\delta}} \right]^{\delta} \right) \quad (4)$$

Avec :

- β_1 et β_2 les paramètres de **forme**,
- θ_1 et θ_2 les paramètres d'**échelle**,
- δ le paramètre de **corrélation**.

Estimation des paramètres

- $\beta_1, \beta_2, \theta_1, \theta_2$: paramètres des lois marginales
- δ est estimé en utilisant la méthode des moments

Les moments de la loi

$$E[X] = \theta_1 \Gamma(1/\beta_1 + 1) \quad (5)$$

$$E[Y] = \theta_2 \Gamma(1/\beta_2 + 1) \quad (6)$$

$$\text{Var}(X) = \theta_1^2 (\Gamma(2/\beta_1 + 1) - \Gamma^2(1/\beta_1 + 1)) \quad (7)$$

$$\text{Var}(Y) = \theta_2^2 (\Gamma(2/\beta_2 + 1) - \Gamma^2(1/\beta_2 + 1)) \quad (8)$$

$$\text{Cov}(X, Y) = \theta_1 \theta_2.$$

$$\begin{aligned} & [\Gamma(\delta/\beta_1 + 1) \Gamma(\delta/\beta_2 + 1) \Gamma(1/\beta_1 + 1/\beta_2 + 1) \\ & - \Gamma(1/\beta_1 + 1) \Gamma(1/\beta_2 + 1) \Gamma(\delta/\beta_1 + \delta/\beta_2 + 1)] \quad (9) \\ & \div \Gamma(\delta/\beta_1 + \delta/\beta_2 + 1) \end{aligned}$$

L'estimation de δ Lachambre, Obrecht, Pinquier 2009

Étape 1 : M.q. δ est solution d'une équation

$$\text{Cov}(X, Y) = \dots \Leftrightarrow \delta B(\delta/\beta_1, \delta/\beta_2) = C = f(\delta)$$

Avec $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ la fonction Beta

L'estimation de δ Lachambre, Obrecht, Pinquier 2009

Étape 1 : M.q. δ est solution d'une équation

$$\text{Cov}(X, Y) = \dots \Leftrightarrow \delta B(\delta/\beta_1, \delta/\beta_2) = C = f(\delta)$$

Étape 2 : M.q. $f(\delta)$ est strictement décroissante – Unicité de δ

$$f'(\delta) = B(\delta/\beta_1, \delta/\beta_2) [1 + \delta/\beta_1 (\psi_0(\delta/\beta_1) - \psi_0(\delta/\beta_1 + \delta/\beta_2)) + \delta/\beta_2 (\psi_0(\delta/\beta_2) - \psi_0(\delta/\beta_1 + \delta/\beta_2))]$$

Avec $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ la fonction Beta, $\psi_0(x) = \frac{d \ln \Gamma(x)}{dx}$ la fonction digamma

L'estimation de δ Lachambre, Obrecht, Pinquier 2009

Étape 1 : M.q. δ est solution d'une équation

$$\text{Cov}(X, Y) = \dots \Leftrightarrow \delta B(\delta/\beta_1, \delta/\beta_2) = C = f(\delta)$$

Étape 2 : M.q. $f(\delta)$ est strictement décroissante – Unicité de δ

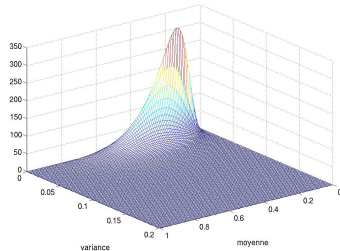
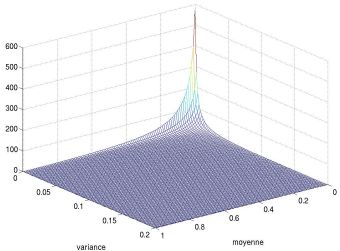
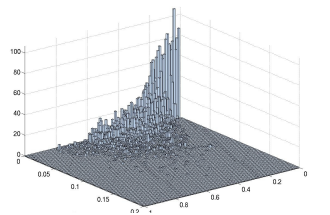
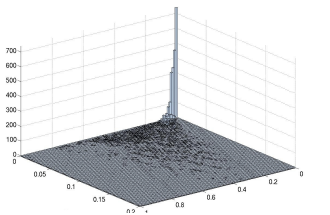
$$f'(\delta) = B(\delta/\beta_1, \delta/\beta_2) [1 + \delta/\beta_1 (\psi_0(\delta/\beta_1) - \psi_0(\delta/\beta_1 + \delta/\beta_2)) + \delta/\beta_2 (\psi_0(\delta/\beta_2) - \psi_0(\delta/\beta_1 + \delta/\beta_2))]$$

Étape 3

$$a(\psi_0(a) - \psi_0(a + b)) + b(\psi_0(b) - \psi_0(a + b)) < -1 \Rightarrow f'(\delta) < 0$$

Avec $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ la fonction Beta, $\psi_0(x) = \frac{d \ln \Gamma(x)}{dx}$ la fonction digamma et $a = \delta/\beta_1$ et $b = \delta/\beta_2$, $(a, b) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*}$

Distributions de Weibull estimées – Validation visuelle



(a) Monophonie

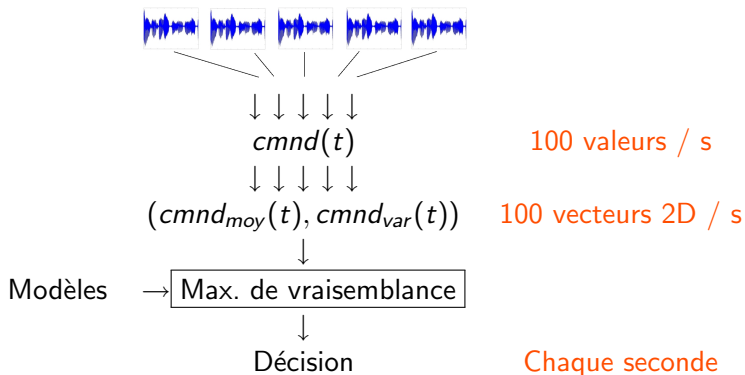
(b) Polyphonie

Validation théorique des marginales

TAB.: Test de Kolmogorov

		Weibull	Gaussienne	Gamma
Poly	$cmnd_{moy}$	0.0642 ✓	0.0746 ✓	0.0757 ✓
	$cmnd_{var}$	0.0863 ✓	0.235 ✗	0.107 ✗
Mono	$cmnd_{moy}$	0.092 ✓	0.228 ✗	0.082 ✓
	$cmnd_{var}$	0.335 ✗	0.274 ✗	0.358 ✗

Fonctionnement du système



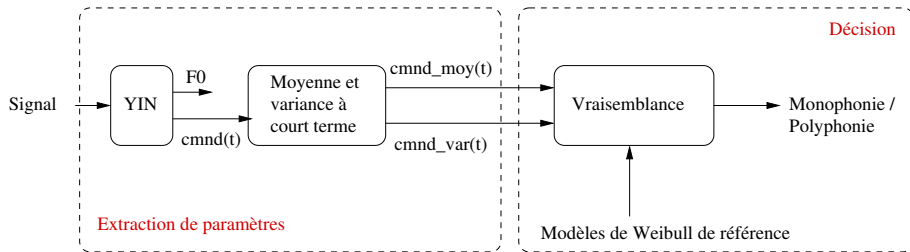
Approche « Classe »

Test de l'approche proposée :

● Approche « Classe » : *cmnd* + Weibull bivariées

Comparaisons et validation de la méthode :

- Système « de base » : 26 MFCC + GMM (16 composantes)
- Système « Gaussien » : *cmnd* + Gaussiennes bivariées
- Système « Weibull univarié » : *cmnd* + Weibull univariées
- Système « SVM » : *cmnd* + SVM (noyau gaussien)



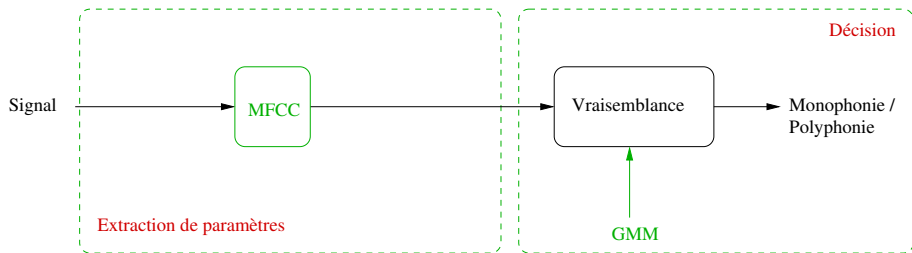
Système « de Base » – Référence

Test de l'approche proposée :

- Approche « Classe » : *cmd* + Weibull bivariées

Comparaisons et validation de la méthode :

- **Système « de base »** : 26 MFCC + GMM (16 composantes)
- Système « Gaussien » : *cmd* + Gaussiennes bivariées
- Système « Weibull univarié » : *cmd* + Weibull univariées
- Système « SVM » : *cmd* + SVM (noyau gaussien)



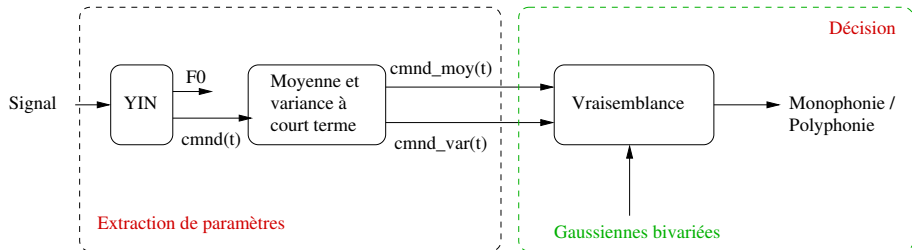
Système « Gaussien » – Validation des paramètres et de la modélisation

Test de l'approche proposée :

- Approche « Classe » : *cmnd* + Weibull bivariées

Comparaisons et validation de la méthode :

- Système « de base » : 26 MFCC + GMM (16 composantes)
- **Système « Gaussien »** : *cmnd* + Gaussiennes bivariées
- Système « Weibull univarié » : *cmnd* + Weibull univariées
- Système « SVM » : *cmnd* + SVM (noyau gaussien)



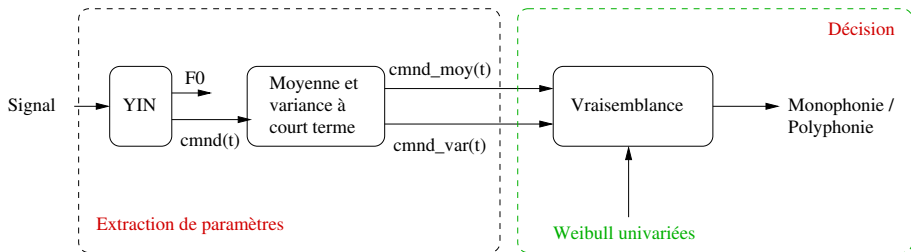
Système « Weibull univarié » – Validation de l'approche **bivariée**

Test de l'approche proposée :

- Approche « Classe » : *cmnd* + Weibull bivariées

Comparaisons et validation de la méthode :

- Système « de base » : 26 MFCC + GMM (16 composantes)
- Système « Gaussien » : *cmnd* + Gaussiennes bivariées
- **Système « Weibull univarié »** : *cmnd* + Weibull univariées
- Système « SVM » : *cmnd* + SVM (noyau gaussien)



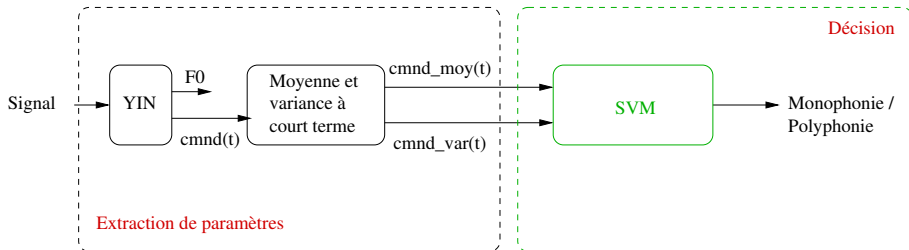
Système « SVM » – Validation de l'approche **probabiliste**

Test de l'approche proposée :

- Approche « Classe » : *cmnd* + Weibull bivariées

Comparaisons et validation de la méthode :

- Système « de base » : 26 MFCC + GMM (16 composantes)
- Système « Gaussien » : *cmnd* + Gaussiennes bivariées
- Système « Weibull univarié » : *cmnd* + Weibull univariées
- **Système « SVM »** : *cmnd* + SVM (noyau gaussien)



Résultats – Comparaison

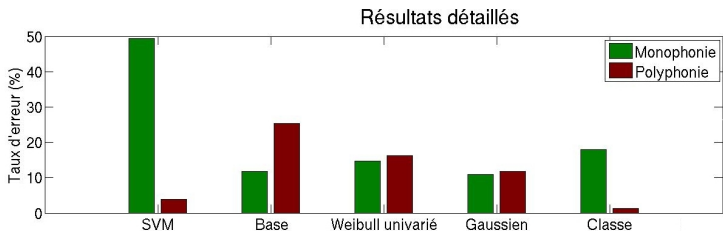
TAB.: Taux global d'erreur de chaque approche.

SVM	Base	W. univ.	Gauss.	Classe
22,5 %	19,2 %	15,5 %	11,4 %	8,5 %

Résultats – Comparaison

TAB.: Taux global d'erreur de chaque approche.

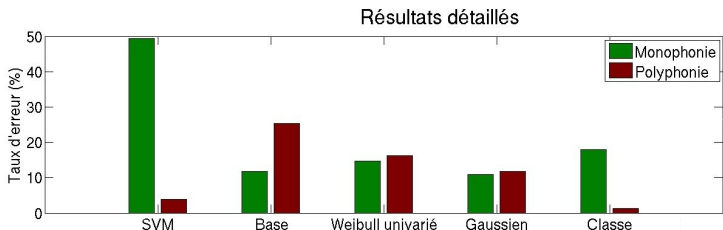
SVM	Base	W. univ.	Gauss.	Classe
22,5 %	19,2 %	15,5 %	11,4 %	8,5 %



Résultats – Comparaison

TAB.: Taux global d'erreur de chaque approche.

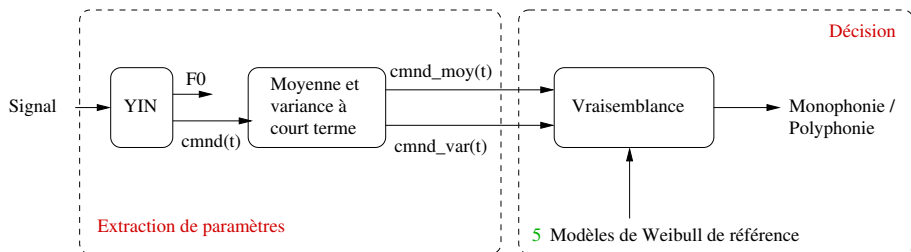
SVM	Base	W. univ.	Gauss.	Classe
22,5 %	19,2 %	15,5 %	11,4 %	8,5 %



Erreurs :

- Notes monophoniques rapides
- Accords polyphoniques « trop parfaits »

Approche « Sous-Classe » : 5 modèles de Weibull bivariés



Taux global d'erreur = 6,5 % (gain de 2 %)

Outline

- ① Problématique
- ② Monophonique / Polyphonique
- ③ **Détection du chant**
 - État de l'art
 - Notre approche
 - Paramètres
 - Décision
 - Expériences et résultats
- ④ Conclusion

État de l'art – Paramètres & Modélisation (1/2)

- MFCC [*Rocamora et al. 2007, Markaki et al. 2008*]
- Paramètres spectraux (centroïde, flux, énergie par bande de fréquence, LPCC) [*Rocamora et al. 2007, Ramona et al. 2008, Berenzweig et al. 2001*]
- Paramètres temporels : ZCR, Énergie, Ondelettes [*Rocamora et al. 2007, Ramona et al. 2008, Zang 2003, Mesaros et al. 2006*]

État de l'art – Paramètres & Modélisation (2/2)

- Paramètres spécifiques :
 - Coefficient harmonique [Chu et al. 2001]
 - Analyse de l'évolution temporelle des fréquences :
modulation d'amplitude dans les bandes de fréquences,
stabilité des harmoniques, trajectoire de F_0 [Maddage 2004, Ohishi
2005, Santosh et al. 2009]
 - Vibrato [Gerhard 2002, Regnier et al. 2009, Khine et al. 2008, Nwe et al. 2007]
- Modélisation : GMM [Lukashevich et al. 2007, Toh et al. 2008, Ezzaidi et al. 2002],
SVM [Ramona et al. 2008, Rocamora et al. 2007], HMM [Berenzweig et al. 2001, Ramona et al.
2008]

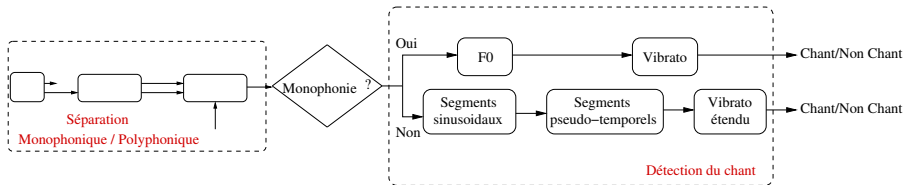
État de l'art – Corpora et Résultats

Pas de corpus de référence :

- Sons monophoniques [*Gerhard 2002, Ohishi 2005*]
- Chant dans la musique Pop [*Maddage 2004, Nwe 2004, Khine et al. 2008, Rocamora et al. 2007*]
- Enregistrements « faits maison » [*Mesaros et al. 2006, Ohishi 2005*]
- Quelques corpora très variés [*Taniguchi 2005, Ezzaidi et al. 2002*]

Résultats : 75 % à 80 % de bonne détection

Schéma global

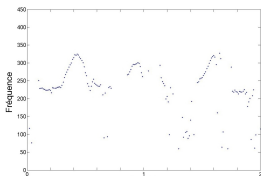


Vibrato

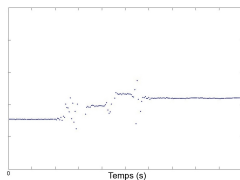
Vibrato

Oscillation périodique de la fréquence fondamentale

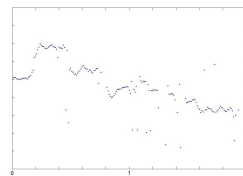
Vibrato du chant : **toujours présent** et à une fréquence **entre 4 et 8 Hz**



(a) Parole



(b) Instrument



(c) Chant

FIG.: Fréquence fondamentale (extraits de 2 s).

Segments sinusoïdaux : Suivi des maxima du spectre [Taniguchi et al. 2005]

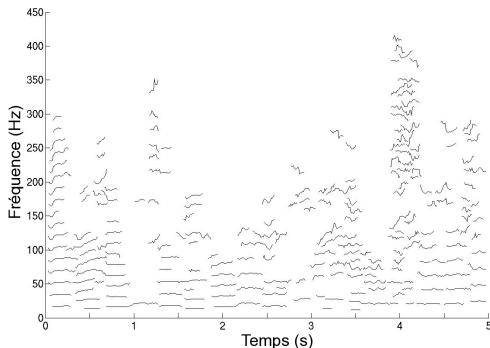


FIG.: Chanson *a capella* (5 s) : chaque ligne est un segment sinusoïdal.

Un segment sinusoïdal est défini par :

- un instant de début
- un instant de fin
- un vecteur de fréquences

Segments sinusoïdaux : Suivi des maxima du spectre [Taniguchi et al. 2005]

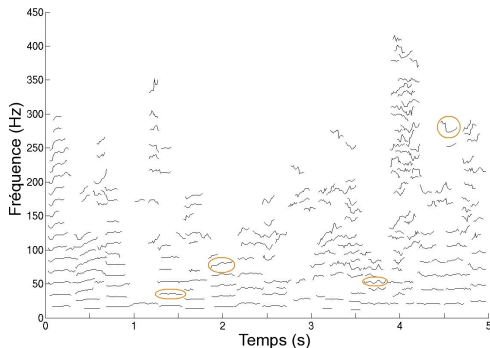


FIG.: Chanson *a capella* (5 s) : chaque ligne est un segment sinusoïdal.
Un segment sinusoïdal est défini par :

- un instant de début
- un instant de fin
- un vecteur de fréquences

Segments sinusoïdaux : Suivi des maxima du spectre [Taniguchi et al. 2005]

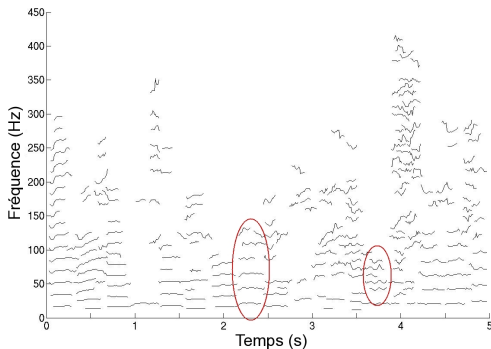


FIG.: Chanson *a capella* (5 s) : chaque ligne est un segment sinusoïdal. Un segment sinusoïdal est défini par :

- un instant de début
- un instant de fin
- un vecteur de fréquences

Segments pseudo-temporels [Lachambre, André-Obrecht, Pinquier 2006]

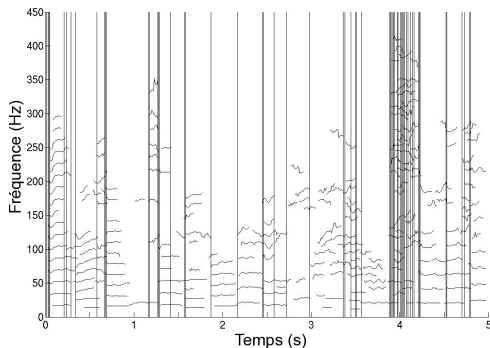


FIG.: Segmentation pseudo-temporelle du même extrait : les lignes verticales sont les limites des segments.

Vibrato étendu [Lachambre, André-Obrecht, Pinquier 2006]

Le vibrato étendu

$$vibr = \frac{\sum_{s \in \Gamma} I(s)}{\sum_{s \in \Omega} I(s)} \quad (10)$$

Avec :

- Ω les segments sinusoïdaux présents dans le segment pseudo-temporel courant
- Γ les segments sinusoïdaux avec du vibrato
- $I(s)$ la durée du segment sinusoïdal s

Décision

Chaque seconde

Monophonie

Il y a du chant si on trouve du **vibrato** sur au moins 10 % des trames.

Polyphonie

Il y a du chant si **vibr** $> 0,15$ sur au moins une trame.

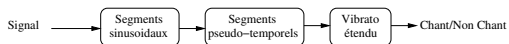
Système « de base »

- **Système « de base »** : GMM + MFCC
- Système « primaire » : vibrato étendu
- Système « manuel » : Segmentation Mono/Poly manuelle, vibrato, vibrato étendu
- Système « automatique » : Segmentation Mono/Poly automatique, vibrato, vibrato étendu



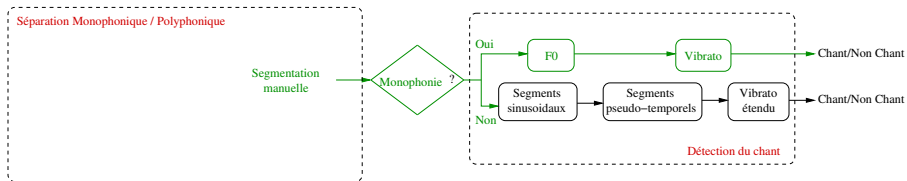
Système « primaire »

- Système « de base » : GMM + MFCC
- **Système « primaire » : vibrato étendu**
- Système « manuel » : Segmentation Mono/Poly manuelle, vibrato, vibrato étendu
- Système « automatique » : Segmentation Mono/Poly automatique, vibrato, vibrato étendu



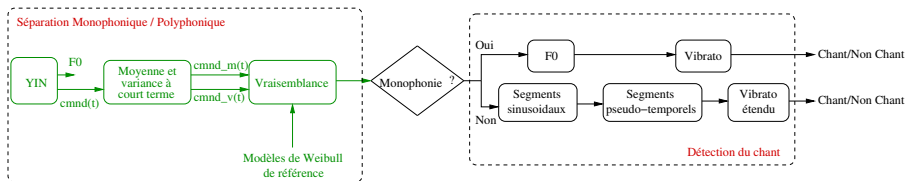
Système « manuel »

- Système « de base » : GMM + MFCC
- Système « primaire » : vibrato étendu
- **Système « manuel »** : Segmentation Mono/Poly manuelle, vibrato, vibrato étendu
- Système « automatique » : Segmentation Mono/Poly automatique, vibrato, vibrato étendu



Système « automatique »

- Système « de base » : GMM + MFCC
- Système « primaire » : vibrato étendu
- Système « manuel » : Segmentation Mono/Poly manuelle, vibrato, vibrato étendu
- **Système « automatique »** : Segmentation Mono/Poly automatique, vibrato, vibrato étendu



Résultats – Comparaison

TAB.: Taux global d'erreur de chaque approche.

Base	Primaire	Manuel	Automatique
28,7 %	29,7 %	21,7 %	25 %

Résultats détaillés- Segmentation Mono/Poly automatique

TAB.: Résultats - Système automatique.

	Chant	Non chant
Chanteur solo	79 %	21 %
Instrument solo	26 %	74 %
Instruments et chanteurs	65 %	35 %
Instruments	18 %	82 %

Résultats détaillés- Segmentation Mono/Poly automatique

TAB.: Résultats - Système automatique.

	Chant	Non chant
Chanteur solo	79 %	21 %
Instrument solo	26 %	74 %
Instruments et chanteurs	65 %	35 %
Instruments	18 %	82 %

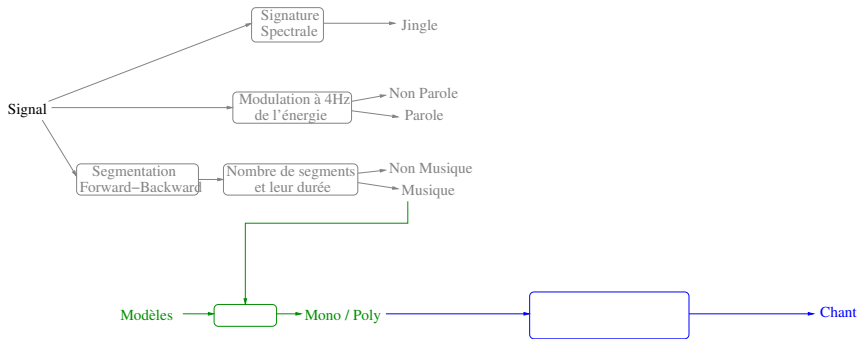
Erreurs :

- Instruments à vent avec du vibrato
- Chant accompagné de nombreux instruments

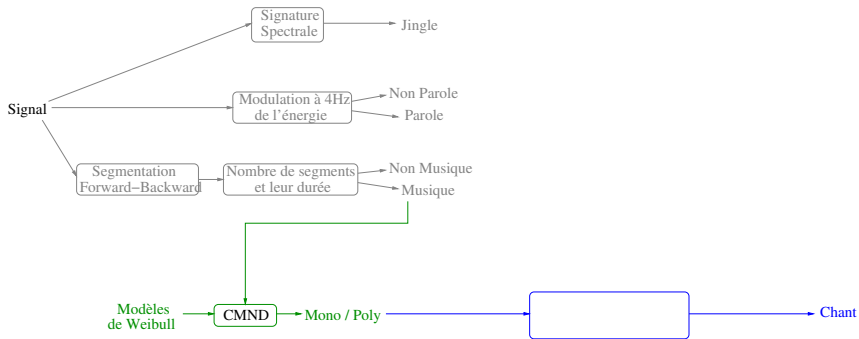
Outline

- 1 Problématique
- 2 Monophonique / Polyphonique
- 3 Détection du chant
- 4 Conclusion**
 - Contribution
 - Synthèse – Structuration
 - Perspectives

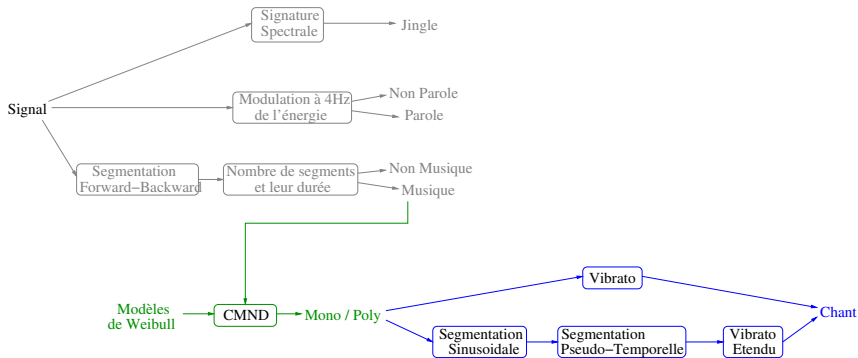
Bilan



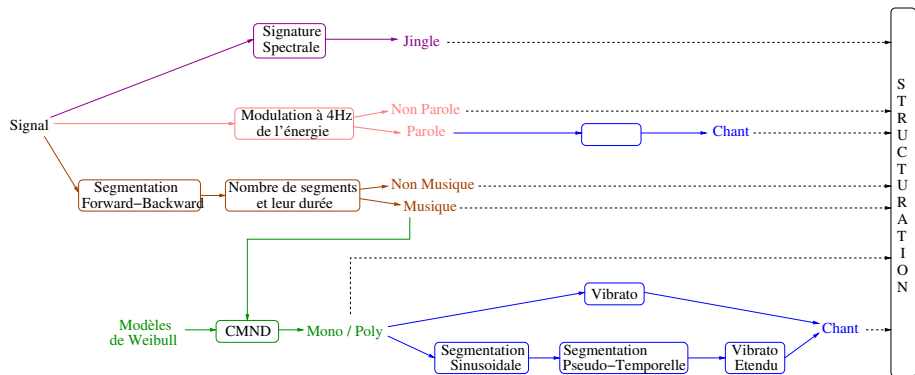
Bilan



Bilan



Une tentative de structuration



Une tentative de structuration – Exemple 1

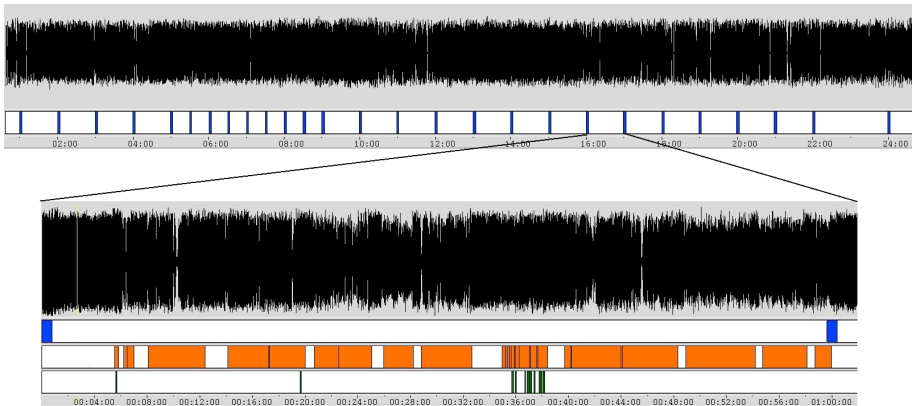


FIG.: France Inter – 16h00-17h00 : Emission de musique instrumentale

Jingles, Musique, Chant

Une tentative de structuration – Exemple 2

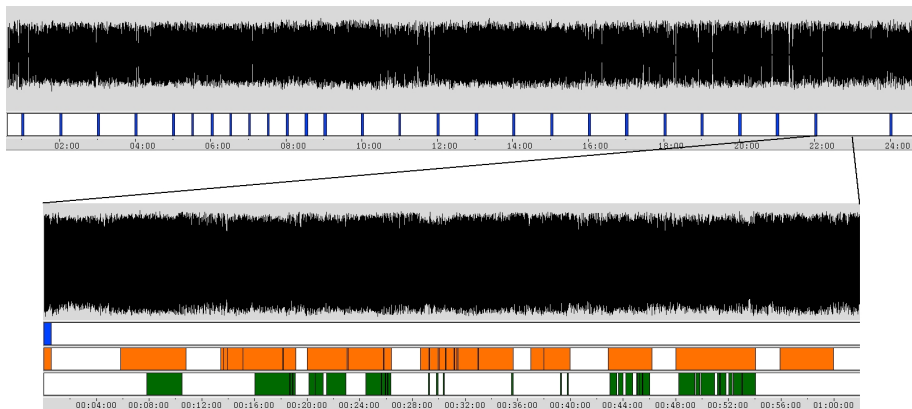


FIG.: France Inter – 22h00-23h00 : Emission musicale – Variétés
 Jingles, Musique, Chant

Perspectives – Monophonique / Polyphonique

TAB.: Test de Kolmogorov

		Weibull	Gaussienne	Gamma
Poly	$cmnd_{moy}$	0.0642 ✓	0.0746 ✓	0.0757 ✓
	$cmnd_{var}$	0.0863 ✓	0.235 ✗	0.107 ✗
Mono	$cmnd_{moy}$	0.092 ✓	0.228 ✗	0.082 ✓
	$cmnd_{var}$	0.335 ✗	0.274 ✗	0.358 ✗

⇒ Utiliser des lois différentes ?

Perspectives – Monophonique / Polyphonique

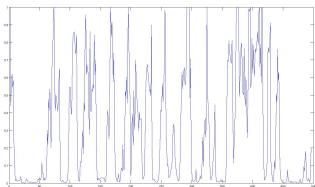
Application à la distinction Parole / Parole Simultanée

- Traduction simultanée
- Personnes se coupant la parole → zones très courtes

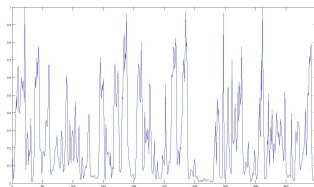
Perspectives – Monophonique / Polyphonique

Application à la distinction Parole / Parole Simultanée

- Traduction simultanée
- Personnes se coupant la parole → zones très courtes



(a) Parole



(b) Parole simultanée

FIG.: Valeurs de $cmd(t)$ pour 5 secondes de signal.

Perspectives – Chant

Amélioration de notre méthode :

- Fusion avec le système de base
- Ajout de paramètres : Coefficient harmonique, paramètres spectraux, MFCC... → Sélection de paramètres

Questions ?

SVM

