IRISA

institut de recherche en informatique
et systèmes aléatoires

france telecom

One Microphone Singing Voice Separation using Source – Adapted Models

WASPAA 17 October 2005

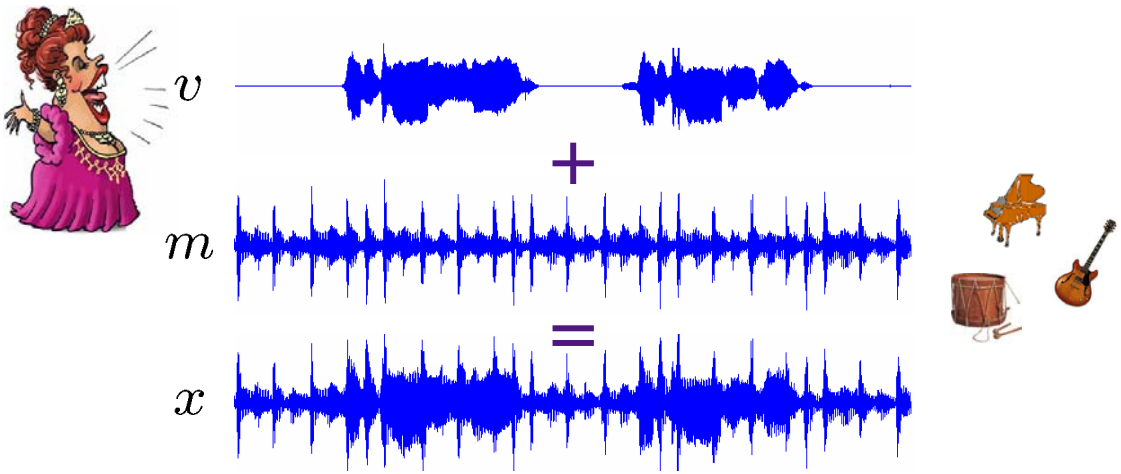
Alexey OZEROV, Pierrick PHILIPPE (FTR&D),

Rémi GRIBONVAL, Frédéric BIMBOT (IRISA – METISS project)

Outline

- Introduction
- GMM – Based One Microphone Source Separation
- Model Adaptation
- Experimentations and Results
- Conclusions and Further Work

Introduction



$$x(n) = v(n) + m(n)$$

mixture **voice** **music**

No spatial information

Cannot use ICA

Need for another *a priori* knowledge

Probabilistic models

$x(n)$ **is observed**

Aim: estimate the voice contribution $\hat{v}(n)$

Application: voice pitch analysis, lyrics recognition, etc.

Outline

- Introduction
- GMM – Based One Microphone Source Separation
  - Source Modeling
  - Model Learning
  - Source Estimation
  - How does it work?
- Model Adaptation
- Experimentations and Results
- Conclusions and Further Work

Source Modeling

$$x(n) = v(n) + m(n) \xrightarrow[\text{(STFT)}]{\text{Short Time Fourier Transformation}} X_t(f) = V_t(f) + M_t(f)$$

In the (STFT) domain the sources (voice and music) are modeled by Gaussian Mixture Models (GMMs)

Ephraim 92 [1]
Benaroya 03 [2]

Probability Density Functions (PDFs) of voice and music short time spectra V_t and M_t :

$$p(V_t) = \sum_i \omega_{v,i} N\left(V_t; \bar{0}, \Sigma_{v,i}\right)$$

Local Power Spectral Density (PSD) →

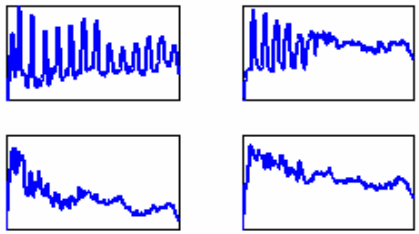
$$\Sigma_{v,i} = \begin{pmatrix} \sigma^2_{v,i}(1) & 0 & \dots & 0 \\ 0 & \sigma^2_{v,i}(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{v,i}(F) \end{pmatrix}$$

$$p(M_t) = \sum_j \omega_{m,j} N\left(M_t; \bar{0}, \Sigma_{m,j}\right)$$

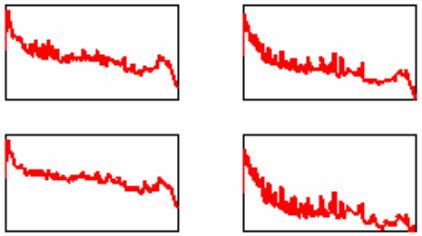
$$\Sigma_{m,j} = \begin{pmatrix} \sigma^2_{m,j}(1) & 0 & \dots & 0 \\ 0 & \sigma^2_{m,j}(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{m,j}(F) \end{pmatrix}$$

← **Local PSD**

Voice GMM



$$\lambda_v = \left\{\omega_{v,i}, \Sigma_{v,i}\right\}_i$$



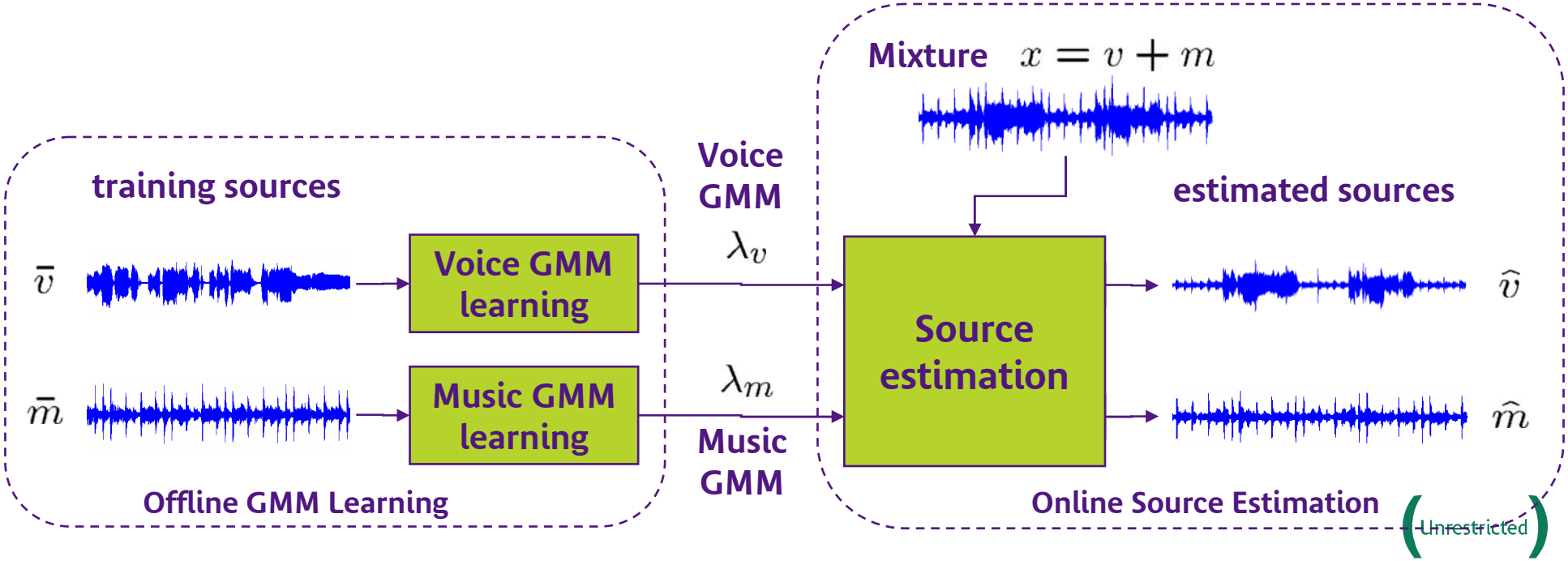
Music GMM

$$\lambda_m = \left\{\omega_{m,j}, \Sigma_{m,j}\right\}_j$$

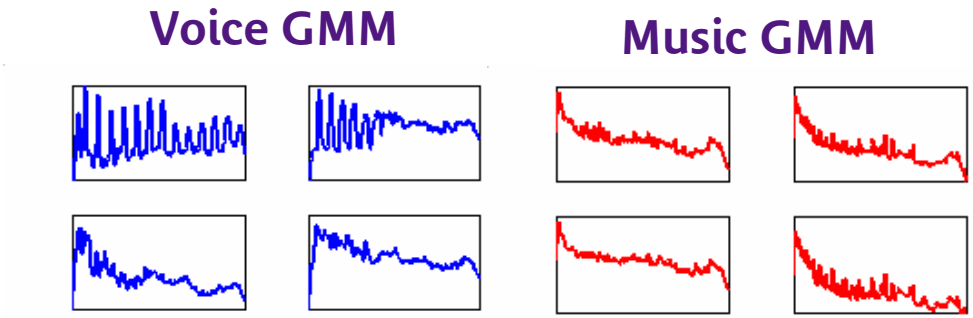
(unrestricted)

Model Learning

- **The models are learned on training sources**
  - Using the Maximum Likelihood (ML) criterion
- **In practice the Expectation – Maximization (EM) Dempster *et al.* 77 [3] algorithm is applied**



Source Estimation



Minimum Mean Square estimator:

Adaptive Wiener filter:

$$\widehat{V}_t(f) = \sum_i \sum_j \underbrace{P\left(q_{v,t} = i, q_{m,t} = j \middle| X, \lambda_v, \lambda_m\right)}_{\text{State probability}} \underbrace{\frac{\sigma^2_{v,i}(f)}{\sigma^2_{v,i}(f) + \sigma^2_{m,j}(f)}}_{\text{Wiener filter}} X_t(f)$$

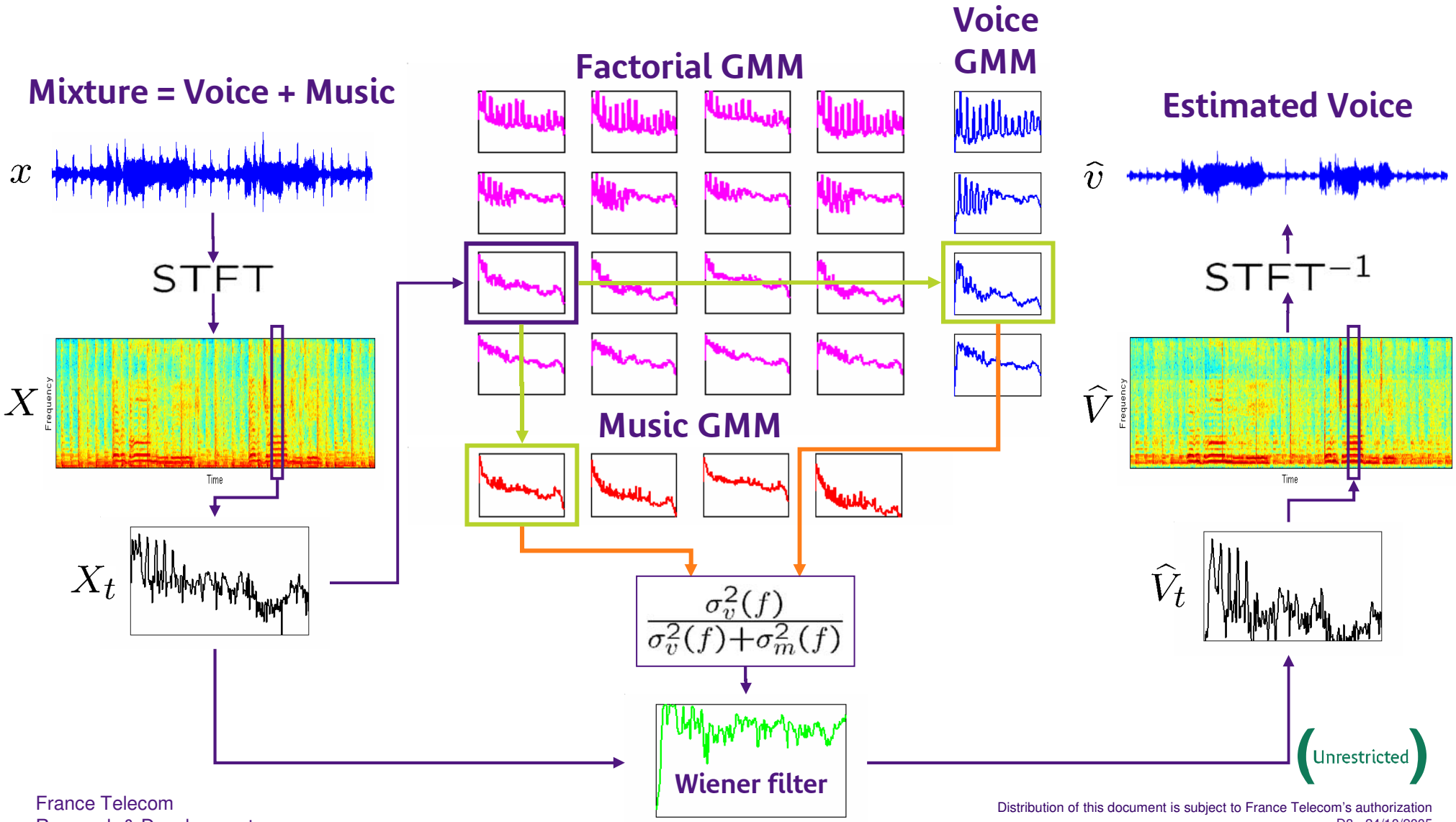
$$\widehat{V}_t(f) \approx \frac{\sigma^2_{v,i^*(t)}(f)}{\sigma^2_{v,i^*(t)}(f) + \sigma^2_{m,j^*(t)}(f)} X_t(f)$$

'Best Wiener filter' approximation

$$(i^*(t), j^*(t)) = \arg\max_{(i,j)} P\left(q_{v,t} = i, q_{m,t} = j \middle| X, \lambda_v, \lambda_m\right)$$

GMM – Based One Microphone Source Separation

How does it work?



(Unrestricted)

Outline

- Introduction
- GMM – Based One Microphone Source Separation
- Model Adaptation
  - Why do we need to Adapt Models?
  - How to Adapt?
  - Voice Model Filter Adaptation
  - Filter – Adapted General Voice Model Learning
- Experimentations and Results
- Conclusions and Further Work

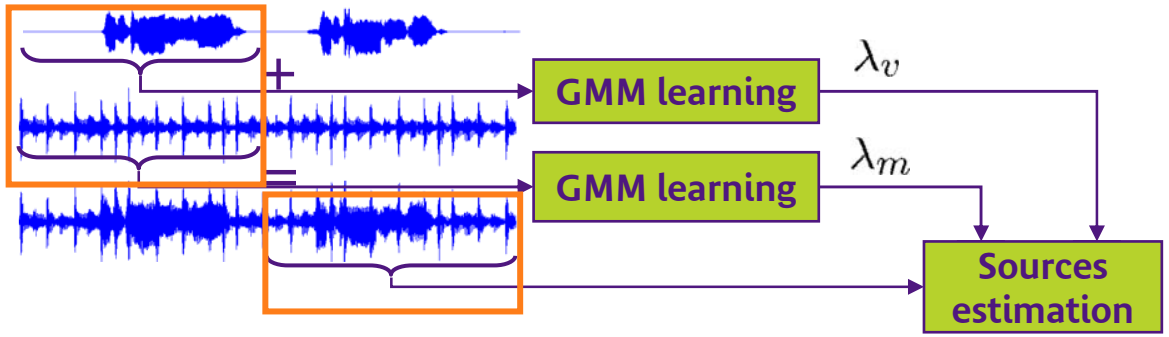
Why do we need to Adapt Models?

What should be used as training sources?

Unrealistic use

Benaroya 03 [2]

the performance is satisfactory



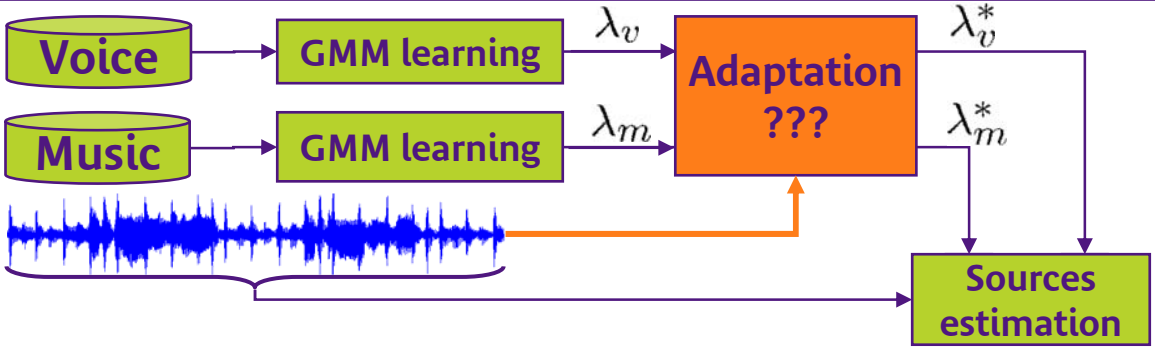
We cannot describe all music and voice variability by some PSDs → **We need to adapt models**

Realistic use

~~General models~~

Adapted models

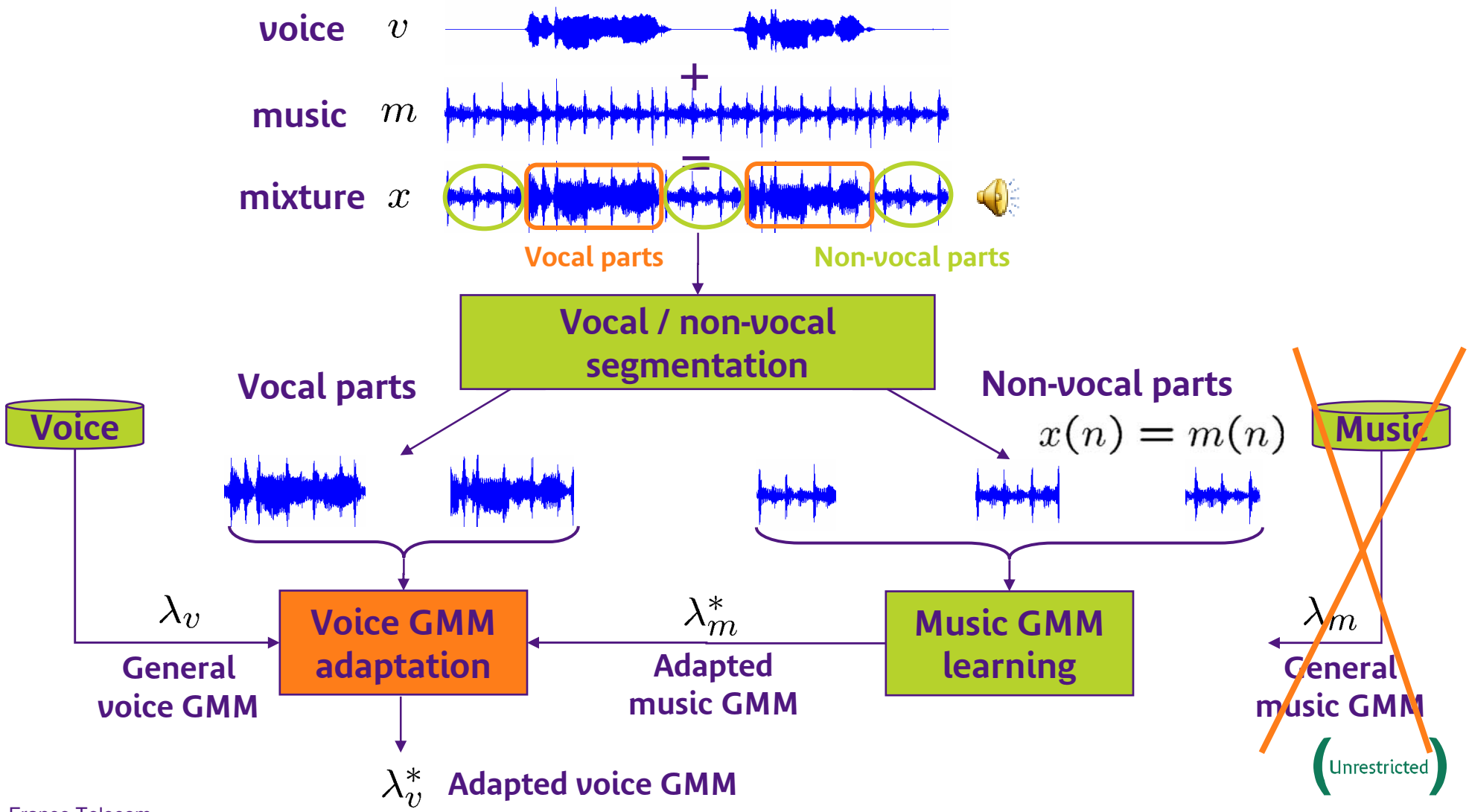
poor performance



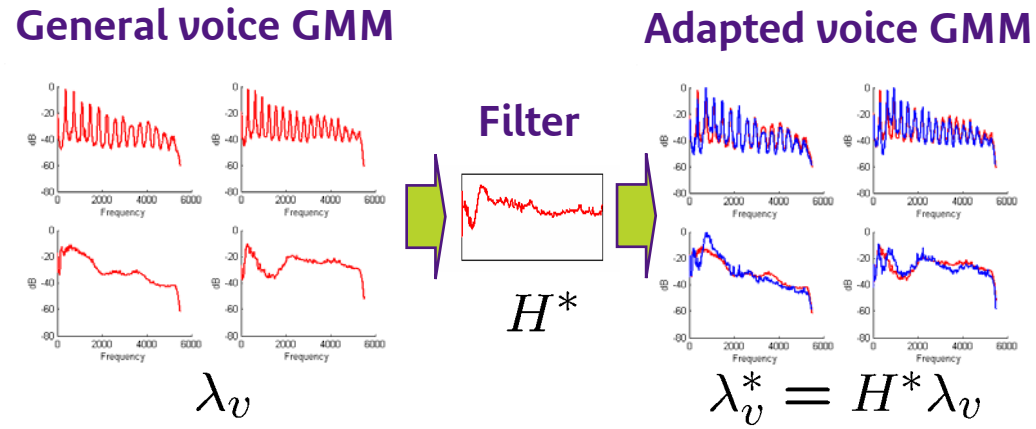
(Unrestricted)

Model Adaptation

How to Adapt?



Voice Model Filter Adaptation



Looking for a filter matching the best the recording of interest in the Maximum Likelihood sense

ML criterion used:

$$H^* = \arg\max_H p(X|H\lambda_v, \lambda_m^*)$$

A priori adapted music GMM

General voice GMM

Filter

$$H = \begin{pmatrix} H(1) & 0 & \dots & 0 \\ 0 & H(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & H(F) \end{pmatrix}$$

This filter adaptation technique makes our modeling invariant to any type of convolutive distortion (room acoustic effects, some microphone characteristics etc …)

Voice Model Filter Adaptation

ML criterion:

$$H^* = \arg\max_H p(X|H\lambda_v, \lambda_m^*)$$

The EM algorithm with following *latent data* is used for optimization:

$$\theta = \{\underbrace{q_v, q_m}_{\text{Discrete model hidden states}}, V\}$$

Voice source

Rose *et al.* 94 [4]

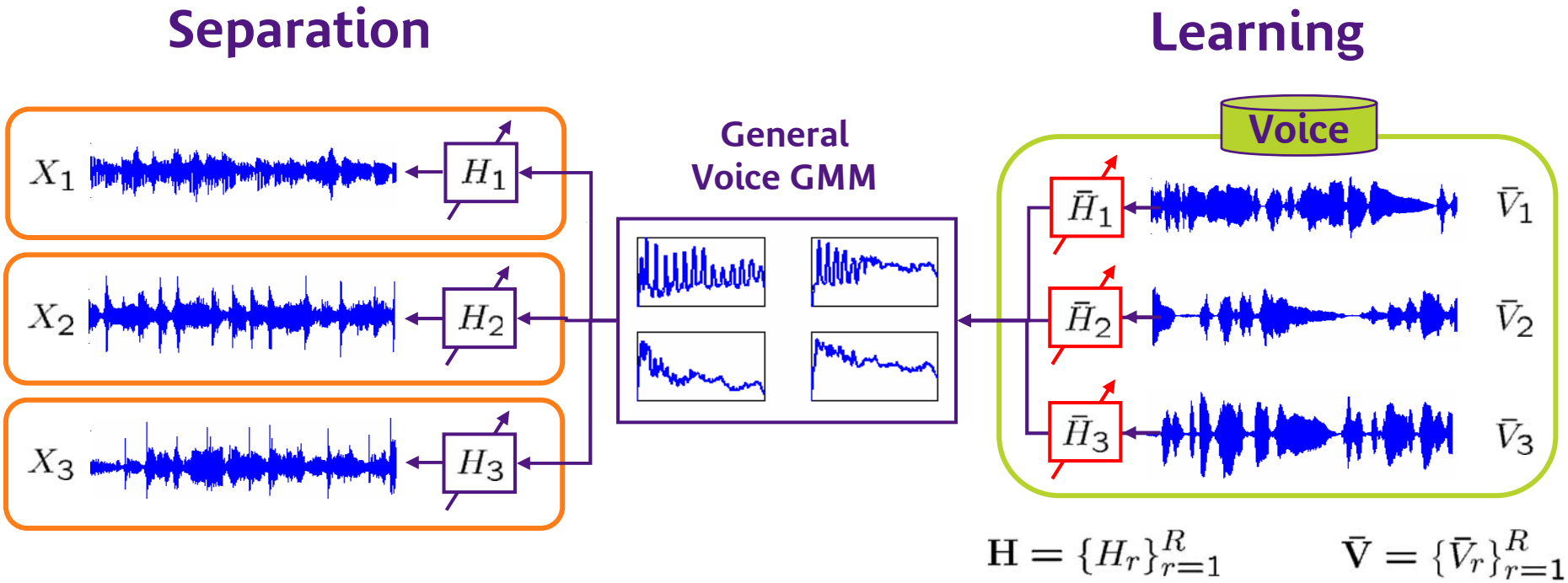
Re-estimation equation:

$$H^{(l+1)}(f) = \frac{1}{T}\sum_t\sum_i \frac{\sum_j \gamma_{i,j}^{(l)}(t)\left\langle |V_t(f)|^2\right\rangle_{i,j}^{(l)}}{\sigma_{v,i}^2(f)}$$

$$\left\langle |V_t(f)|^2\right\rangle_{i,j}^{(l)} \triangleq E\left[|V_t(f)|^2 \middle| X, q_{v,t}=i, q_{m,t}=j, H^{(l)}\lambda_v, \lambda_m^*\right],$$

$$\gamma_{i,j}^{(l)}(t) \triangleq P\left(q_{v,t}=i, q_{m,t}=j \middle| X, H^{(l)}\lambda_v, \lambda_m^*\right)$$

Filter – Adapted General Voice Model Learning



Conventional ML criterion:

$$\hat{\lambda}_v = \arg\max_{\lambda_v} \prod_r p(\bar{V}_r|\lambda_v)$$

Filter - adapted ML criterion:

$$(\hat{\lambda}_v, \mathbf{H}^*) = \arg\max_{(\lambda_v, \mathbf{H})} \prod_r p(\bar{V}_r|\bar{H}_r\lambda_v)$$

(Unrestricted)

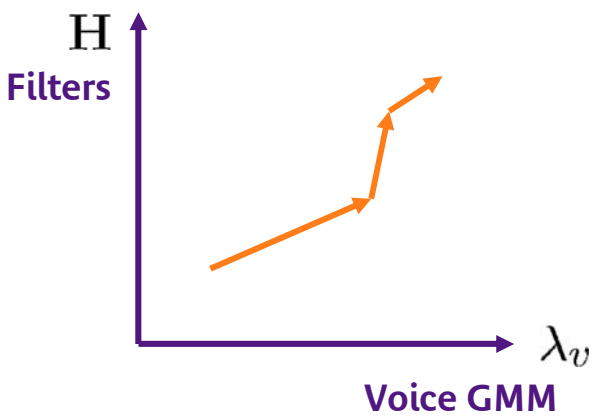
Filter – Adapted General Voice Model Learning

Filter - adapted ML criterion:

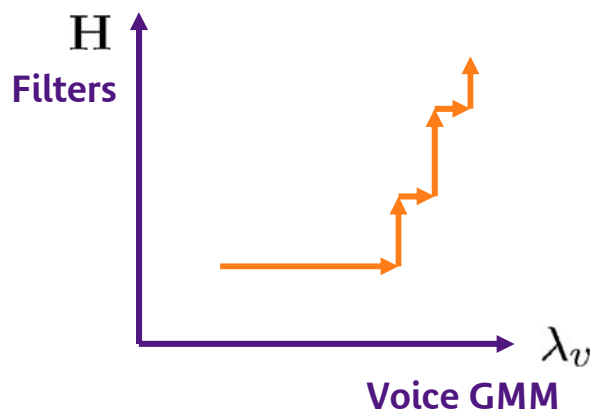
$$(\hat{\lambda}_v, \mathbf{H}^*) = \arg \max_{(\lambda_v, \mathbf{H})} \prod_r p(\bar{V}_r | \bar{H}_r \lambda_v)$$

PROBLEM: it is difficult to solve the M - step

Expectation - Maximization (EM)



Space – Alternating Generalized EM (SAGE) Fessler 94 [5]



Outline

- Introduction
- GMM – Based One Microphone Source Separation
- Model Adaptation
- Experimentations and Results
  - Data Description
  - Performance Measure
  - Simulations
  - Some Audio Examples
- Conclusions and Further Work

Data Description

Training database (for general models)

Voice

- **Singing voice:** 34 samples of singing men's voices from popular music (each sample is approximately 1 min. long)

Music

- **Music:** 30 samples of popular music free from voice (each sample is approximately 1 min. long)

Test database

- **5 songs**
  - Voice and music tracks are available separately.
  - Thus it is possible to evaluate the separation performance by comparing the estimated voice with the original one.
- The test items are manually segmented in vocal and non-vocal parts.
  - Although automatic segmentation is also possible.

All the recordings are mono and sampled at 11025 Hz.

Performance Measure

- **Signal to Distortion Ratio (SDR)** Gribonval *et al.* 03 [6]

$$\mathrm{SDR}(\widehat{v}, v) = 10 \log_{10} \left[\frac{\langle \widehat{v}, v \rangle^2}{\|\widehat{v}\|^2 \|v\|^2 - \langle \widehat{v}, v \rangle^2} \right]$$

- **Normalized SDR (NSDR),** ***SDR improvement*** **between the non-processed mixture x and the estimated voice $\widehat{v}$**

$$\mathrm{NSDR}(\widehat{v}, v, x) = \mathrm{SDR}(\widehat{v}, v) - \mathrm{SDR}(x, v)$$

$\widehat{v}$ **estimated voice**

v **original voice**

x **mixture**

Experimentations and Results

Simulations

32 – states Voice GMM	32 – states Music GMM	NSDR (dB)
Voice general	Music general	5.06
Voice general	learned from non-vocal parts	9.09
filter adapted from vocal parts + filter – adapt. learning	learned from non-vocal parts	10.05

+ 4 dB

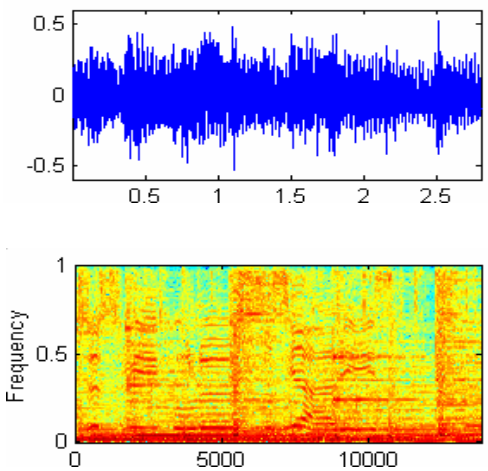
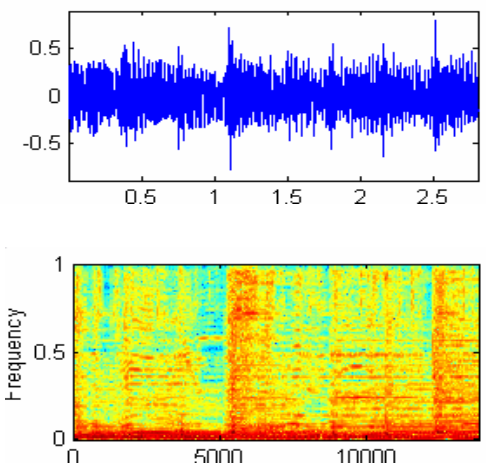
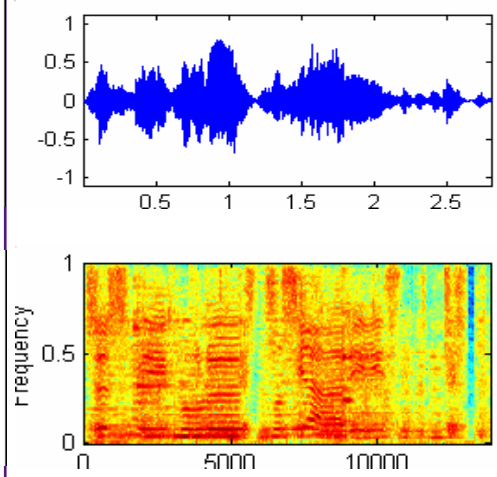
+ 1 dB

Audio Examples !!!

Unrestricted

Experimentations and Results

Some Audio Examples

Mixture x	Estimated music $\hat{m} = x - \hat{v}$	Estimated voice $\hat{v}$
		

Outline

- Introduction
- GMM – Based One Microphone Source Separation
- Model Adaptation
- Experimentations and Results
- Conclusions and Further Work

Conclusions and Further Work

- **Our proposal is based on :**
  - Music model learning on the non - vocal parts
  - Voice model filter adaptation on the vocal parts
  - Filter – adapted general voice model learning
- **5 dB improvement over state of the art**
- **Further work :**
  - Automatic vocal / non – vocal segmentation
  - Joint segmentation / separation

Thank you!

WEB: http://www.irisa.fr/metiss/ozerov/index.html

References

[1] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. SP-40, pp. 725-735, April 1992.

[2] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM / GMM using a single sensor," in *ICA*, Apr 2003.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977.

[4] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identification in noise. *IEEE Trans. Speech and Audio*, 2(2):245–257, April 1994.

References

[5] J. A. Fessler and A. O. Hero. Space-Alternating Generalized Expectation-Maximization algorithm. *IEEE Trans. on Signal Processing*, 42(10), Oct. 1994.

[6] R. Gribonval, L. Benaroya, E. Vincent and C. Févotte, "Proposals for performance measurement in source separation," in *ICA*, Apr 2003, pp. 763 - 768.