Speech and Music Analysis by means of Acoustic Descriptors

Thomas Dubuisson, Thierry Dutoit

TCTS Lab, Faculté Polytechnique de Mons, Belgium

thomas.dubuisson@fpms.ac.be

POLYTECH.MONS

Definition and Context

- Acoustic Descriptor = numerical value for describing an acoustic property of the signal (speech or music).
- Acoustic Descriptor are used to characterize the signal by a limited set of values and to extract information from it.
- Different types of acoustic descriptors exist, distinguished according to 4 points of view:
  1. Steadiness or dynamicity: value extracted from the signal at a given time or a parameter from a model of the signal behavior along time (ex: mean, distribution of a parameter).
  2. Time extent of the description provided by the descriptor: some apply to the whole signal (Global Descriptor) or to a part of it (Local Descriptor).
  3. Abstractness of the descriptor: what the descriptor represents.
  4. Extraction process of the feature: some descriptors are directly computed on the waveform (ex: zero-crossing rate) or after a transformation of the signal (ex: Fourier Transform). Some other relate to a model of the signal or try to mimic the output of the ear system.

Principles of Computation

Signal →

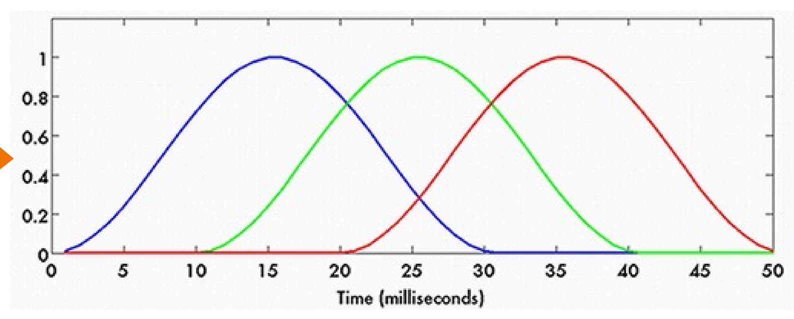


Fig.1 Windowing operation

→ Frame → Temporal Domain ① / Spectral Domain ②

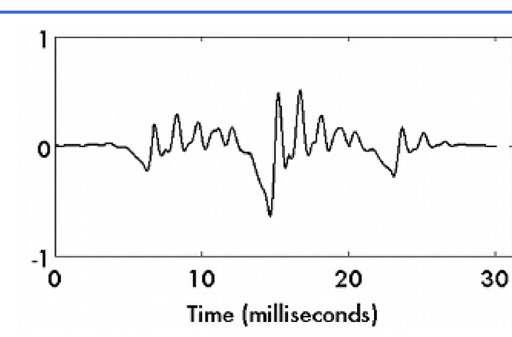
① 

Fig.2 Temporal frame of speech

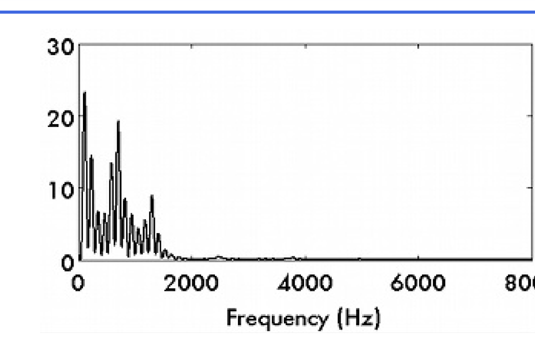
② 

Fig.3 Fourier Transform of the frame of speech

Some Examples of Descriptors

Temporal Domain

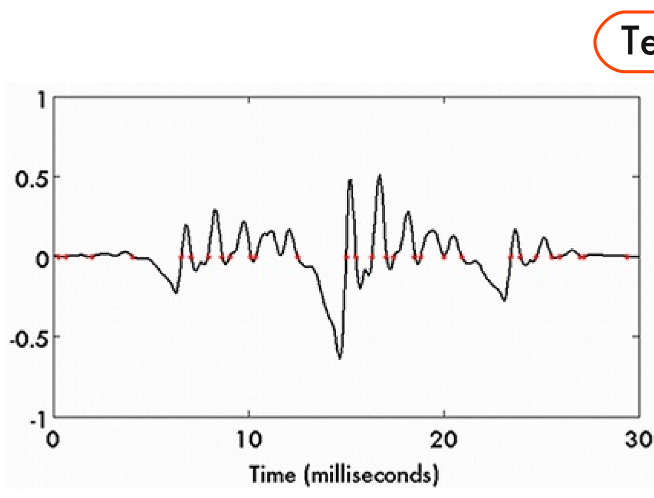


Fig.4 Illustration of the Zero Crossings Detection

$$E_T(dB) = 10 \times log_{10}(\sum_{i=1}^{N} x(n)^2) = 10.10dB$$

$$\mu_T = \frac{1}{N} \times \sum_{i=1}^{N} x(n) = 0$$

$$\sigma_T = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x(n) - \mu_T)^2} = 0.1461$$

$$Zero\ Crossing\ Rate(Hz) = \frac{\#Zero\ Crossing}{0.03} = 1000Hz$$

Spectral Domain

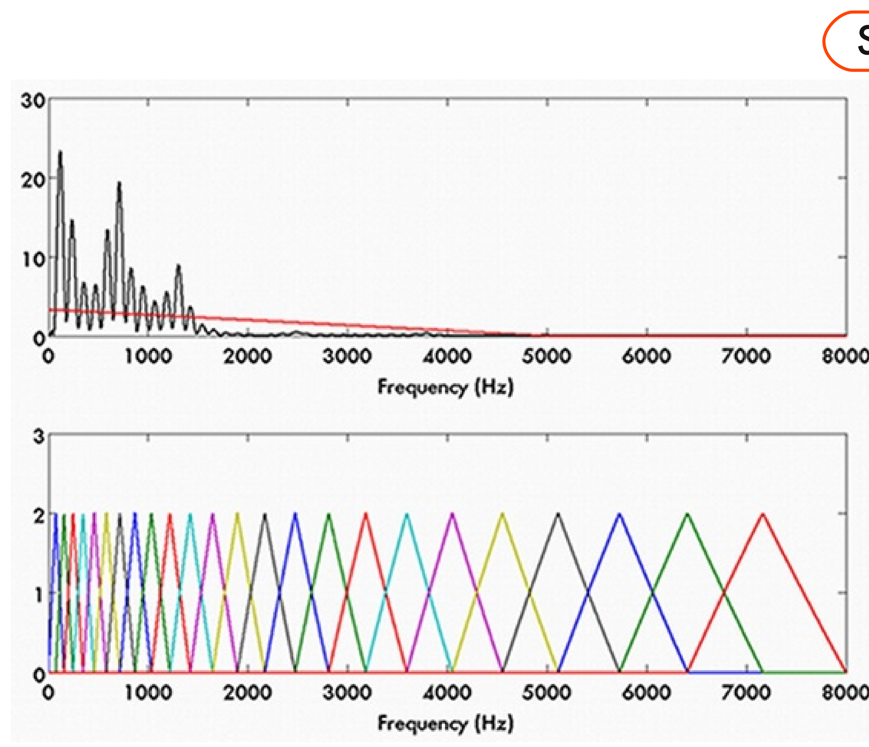


Fig.5 Illustration of the Spectral Slope computation and the MEL scale of frequencies

$$COG(Hz) = \frac{\sum_{f=1}^{8000} f \times X(f)}{\sum_{f=1}^{8000} X(f)} = 749Hz$$

$$Decrease = \frac{\sum_{f=2}^{8000} \frac{X(f) - X(1)}{f-1}}{\sum_{f=2}^{8000} X(f)} = 0.0028$$

$$\hat{X}(f) = Slope \times f + K\ (Slope = -6.62 \times 10^{-4})$$

$$E_0 = \frac{\sum_{f=60}^{f=400} X(f)}{\sum_{f=60}^{f=8000} X(f)} = 0.339$$

$$E_2 = \frac{\sum_{f=2000}^{f=5000} X(f)}{\sum_{f=60}^{f=8000} X(f)} = 0.035$$

$$T_1 = \frac{\sum_{MEL_{[1]}} X(f)}{\sum_{MEL_{[1...24]}} X(f)} = 0.082$$

$$T_2 = \frac{\sum_{MEL_{[2,3,4]}} X(f)}{\sum_{MEL_{[1...24]}} X(f)} = 0.254$$

Applications

Speech Pathologies Analysis

Aim: extracting information from speech signal for finding significant differences between normal speakers and pathological speakers.

Principle: Use of the correlation between 87 acoustic descriptors for discriminating normal and pathological voices.

Database: Kay Elemetrics MEEI Database consisting on 53 normal and 657 pathological sustained vowels /a/.

Computation of the Correlation Matrix

$$R_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) \times (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2} \times \sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

Pearson Coefficient of Correlation

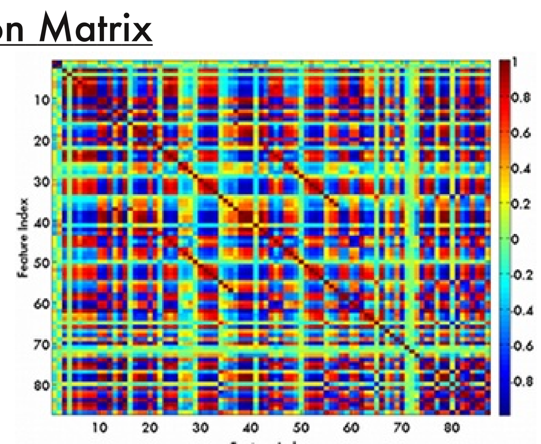


Fig.6 Normophonic Voice

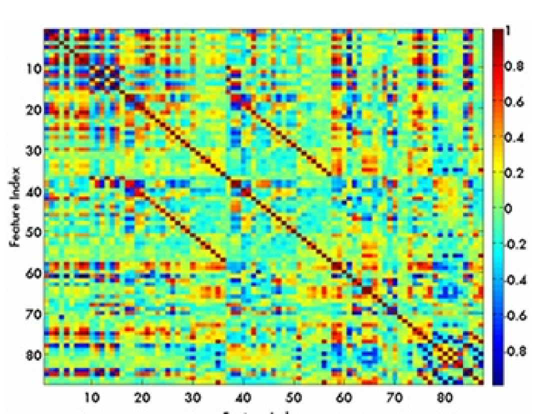


Fig.7 Pathological Voice

Selection of the most discriminant correlation

$$D_k = \frac{\sum_{c=1}^{C} p(\omega_c)(\mu_{ck} - \mu_k)^2}{\sum_{c=1}^{C} p(\omega_c)\sigma_{ck}^2}$$

Discriminant power by Fisher Analysis

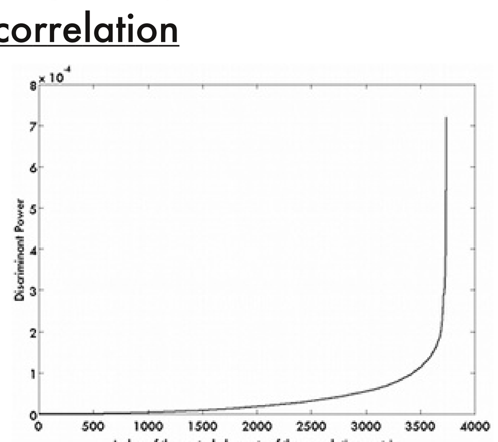


Fig.8 Discriminant power of the correlations (in ascending order)

The correlation between the specral decrease and the first tristimulus in Bark frequency bands is the most discriminant between the two populations.

Discrimination between Normophonic and Pathological Voices

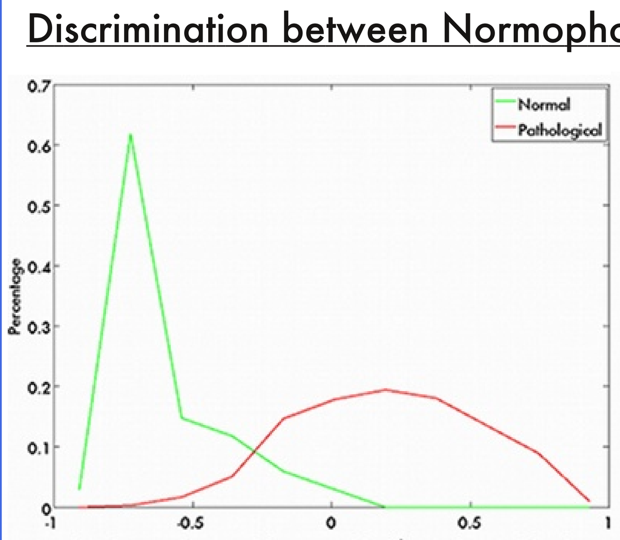


Fig.9 Distribution of the most discriminant correlation for the normal and pathological voices

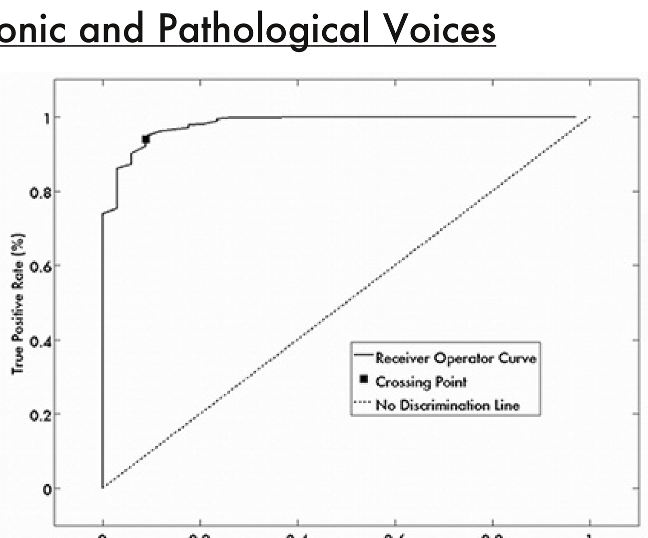


Fig.10 Receiver-Operator Curve for the discrimination between the normal and pathological voices

	Manual Pathological	Manual Normal
Auto Pathological	0.947	0.088
Auto Normal	0.053	0.912

Confusion matrix associated to the crossing point of the distributions

This work is part of the ECLIPSE research project, funded by Région Wallonne (grant WALEO II #516009).

Music Analysis

Aim: extracting information from music signal for browsing through large collection of musical samples or analyzing the structure of a musical excerpt.

Browsing through collection of violin samples

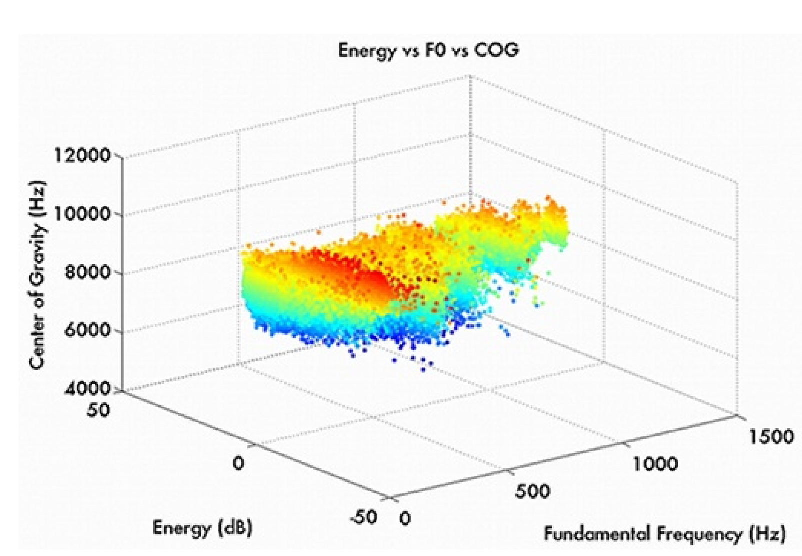


Fig.11 Organization of violin frames in a 3 descriptors space

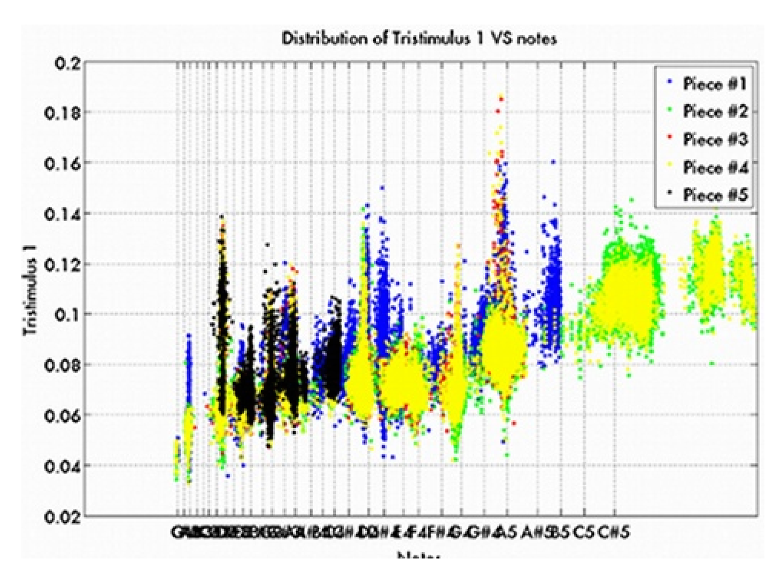


Fig.12 Organization of violin frames in a 2 descriptors space

Analyzing the structure of a musical excerpt

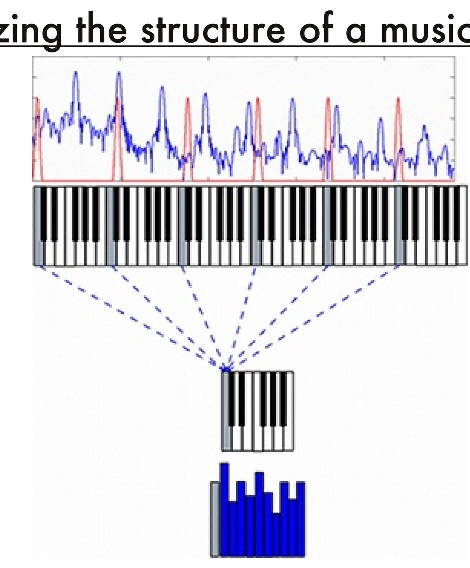


Fig.13 Illustration of the computation of chromatic information

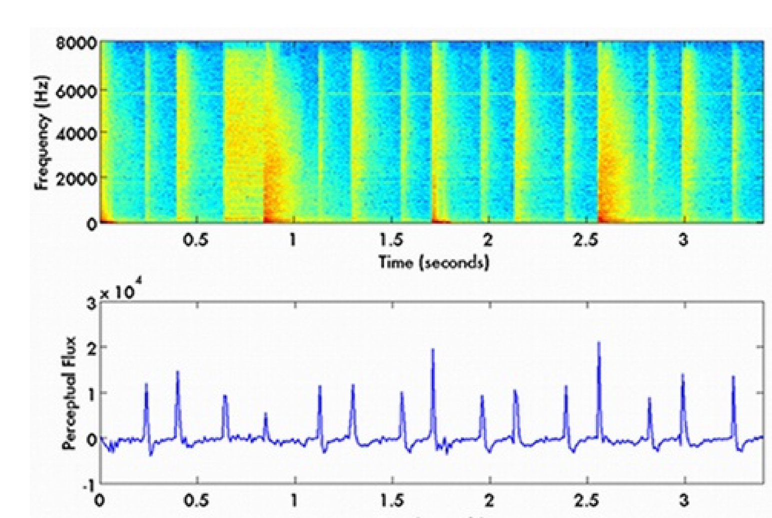
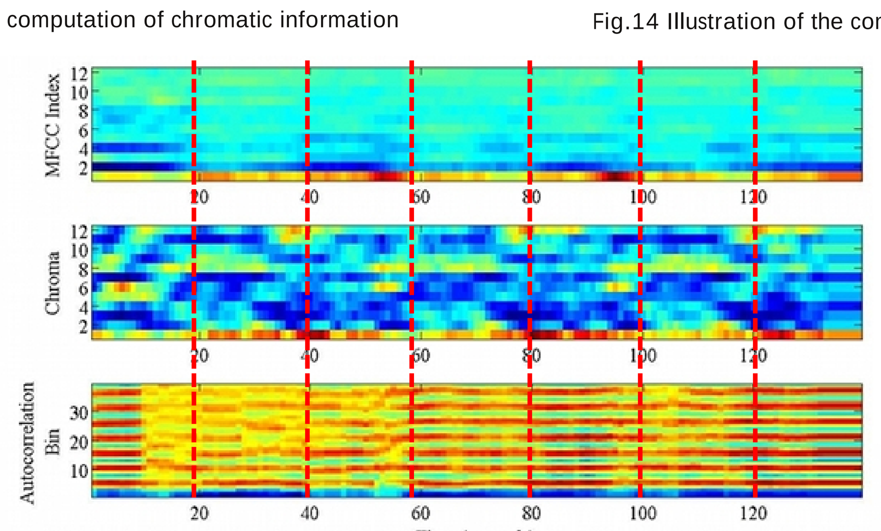


Fig.14 Illustration of the computation of rhytmic information



Timbre analysis

Melody analysis

Rhythm analysis

Fig.15 Analysis of the song 'Foule sentimentale', Alain Souchon

This work is part of the numediart research project, funded by Région Wallonne (grant #716631).

FACULTÉ POLYTECHNIQUE DE MONS

Technologies de l'Information · Région Wallonne · Académie Universitaire Wallonie-Bruxelles