Music Signal Processing and Applications in Recognition

Athanasia Zlatintsi and Petros Maragos

School of ECE, National Technical University of Athens 15773, Greece

http://cvsp.cs.ntua.gr

1. Outline/Contributions

Analysis of music signals for the extraction of effective descriptors for automatic classification.

- Processing/analysis of music signals with nonlinear methods:
  - Fractal theory
  - AM-FM model
- Introduction of new descriptors based on **Bag-of-Words models.**
- Experimental evaluation in applications such as:
  - Recognition of **musical instruments.**
  - Recognition of different **genres of music.**
- Study of the AM-FM model for salient event detection and audio summary creation.
  - Extension of the proposed ideas on multimodal data.
  - Development of a systematic saliency movie database.

2. Motivation

- The power and the role of music in human life from ancient times (Pythagoras, Plato, Aristotle) until today.
- Music's use in entertainment, advertising, cinema, therapy, teaching, work, cultural heritage, etc.
- The amount of digital music and multimedia data in general.

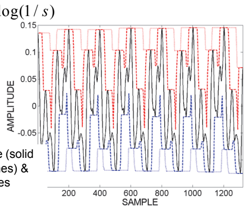
3. Multiscale Fractal Dimension (MFD):

Quantifies the multiscale complexity of the waveform, i.e., the degree of its fragmentation.

- *Fractal Dimension D*

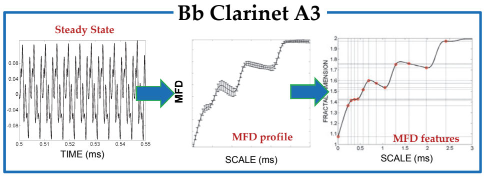
$$D = \lim_{s \to 0} \frac{\log[(\text{Area of dilated graph by disks of radius s}) / s^2]}{\log(1/s)}$$

Algorithm: based on multiscale nonlinear operators of morphological filtering that creates geometrical covers around the graph of the signal.



Bb Clarinet A3 Steady state (solid line) and its dilations (red lines) & erosions (blue lines) at scales s = 25, 75, with D = 1.1.

Bb Clarinet A3



4. Amplitude & Frequency Modulation (AM-FM)

AM-FM model: $s(t) = \sum_{i=1}^{K} \alpha_i(t) \cos(\varphi_i(t))$

instantaneous amplitude — phase

We model each resonance component of music signals as an amplitude and frequency modulated sinusoid (AM-FM signal), and the whole signal as a sum of such AM-FM components.

Energy Separation Algorithm (ESA): estimates the instantaneous amplitude and frequency.

$$|\alpha(t)| \approx \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \qquad f(t) \approx \frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}$$

Teager Energy Operator: $\Psi[x] = \dot{x}^2 - x\ddot{x} \qquad (\dot{x} = dx/dt)$

Gabor-ESA: combination of the continuous time ESA and Gabor filtering of the signal (smoother instantaneous estimates)

$$\Psi[s(t) * g(t)] = \left[s(t) * \frac{dg(t)}{dt} \right]^2 - (s(t) * g(t)) \left[s(t) * \frac{d^2 g(t)}{dt^2} \right]$$

input signal — Gabor impulse response

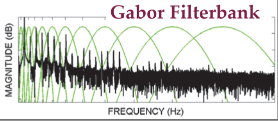
Modulation Features:

Mean Inst. Amplitude

Mean Inst. Frequency

Freq. Modulation Percentage

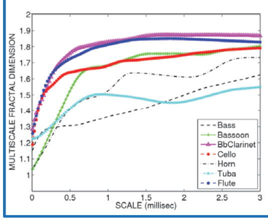
Gabor Filterbank



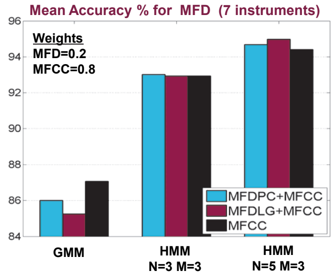
5. Musical Instruments

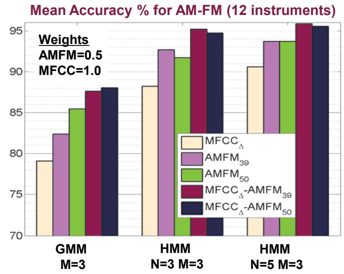
Timbre (instrument specific quality): Quality of sound that distinguishes two sounds of the same pitch, loudness and duration; thus associated with the identification of environmental sound sources.

Example of Average MFDs on Attack of 7 instruments



- Higher D for small scales.
- More fragmented.
- Increased value of D(s = 1).
- Clear distinction of D among some of the instruments.





Conclusions:

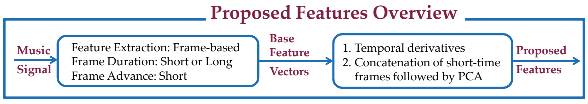
1) Classification with MFD
   - Error Rate Reduction (ERR) up to 32%.
2) Classification with AM-FM
   - ERR up to 38% (AM-FM only).
   - ERR up to 60% for 7 instruments (AM-FM fused with MFCC).
   - ERR up to 56% for 12 instruments (AM-FM fused with MFCC).
3) Iterative-ESA: Possible estimation of the harmonic content of a tone.

6. Musical Genres

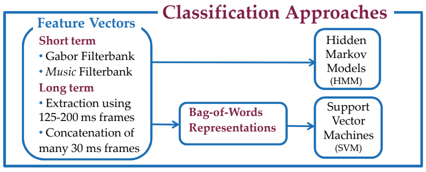
Genre:

- Most widely used term for the description of music; however fuzzy since the boundaries of the various genres are unclear.
- Main method for organizing databases, music stores, music collections etc.
- Recognition problem with great complexity.

Proposed Features Overview

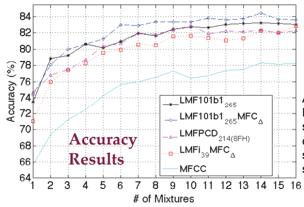


Classification Approaches



Conclusions:

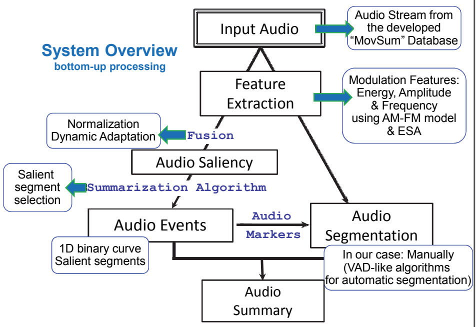
1) **Robust AM-FM Representations:** can capture important aspects of music, such as micro-changes of its structure (e.g., melody, rhythm etc).
1) **"Music" filterbank**
   - Error Rate Reduction (ERR) up to 28%.
2) **Macro-structure Representations** based on the concatenated short-time frames
   - Classification Complexity Reduction: simper statistical models, compact descriptors, shorter training durations.
   - ERR up to 22%.
3) **Bag-of-Words Representations**
   - Compact representations/reduced computational complexity.
   - ERR up to 16% (accuracy 83.6%).



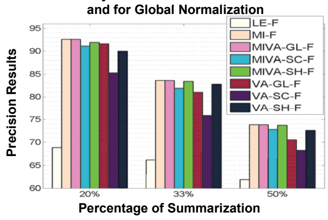
Accuracy results (%) using HMMs for the four best feature sets (proposed feat. only or combined with MFCC for N = 5 states) and for MFCC (N = 7 states) and M = [1,16] mixtures.

7. Audio Summarization

System Overview
bottom-up processing



Summary Evaluation for various Fusions and for Global Normalization



Publications

Journals:

- **A. Zlatintsi** and P. Maragos. **Multiscale Fractal Analysis of Musical Instrument Signals with Application to Recognition**. *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 21(4), pp. 737-748, Apr. 2013.
- G. Evangelopoulos, **A. Zlatintsi**, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas and Y. Avrithis. **Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention**. *IEEE Trans. on Multimedia*, Vol. 15(7), pp. 1553-1568, Nov. 2013.

Conferences:

- **A. Zlatintsi** and P. Maragos. **AM-FM Modulation Features for Music Instrument Signal Analysis and Recognition**. *In Proc. European Signal Process. Conf. (EUSIPCO–12)*, Bucharest, Romania, Aug. 2012.
- **A. Zlatintsi**, P. Maragos, A. Potamianos and G. Evangelopoulos. **A Saliency- Based Approach to Audio Event Detection and Summarization**. *In Proc. European Signal Process. Conf. (EUSIPCO–12)*, Bucharest, Romania, Aug. 2012.
- **A. Zlatintsi** and P. Maragos. **Musical Instruments Signal Analysis and Recognition Using Fractal Features**. *In Proc. European Signal Process. Conf. (EUSIPCO–11)*, Barcelona, Spain, Aug.-Sep. 2011.
- N. Malandrakis and A. Potamianos and G. Evangelopoulos and **A. Zlatintsi. A Supervised Approach to Movie Emotion Tracking**. *In Proc. Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP–11)*, pp. 2376- 2379, Prague, Czech Republic, May 2011.
- G. Evangelopoulos, **A. Zlatintsi**, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos and Y. Avrithis. **Video Event Detection and Summarization Using Audio, Visual and Text Saliency**. *In Proc. Int'l Conf. Acoustics, Speech and Signal Process. (ICASSP–09)*, Taipei, Taiwan, Apr. 2009.
- D. Spachos, **A. Zlatintsi**, V. Moschou, P. Antonopoulos, E. Benetos, M. Kotti, K. Tzimouli, C. Kotropoulos, N. Nikolaidis, P. Maragos and I. Pitas. **MUSCLE Movie Database: A Multimodal Corpus With Rich Annotation For Dialogue And Saliency Detection**. *In Proc. Int'l Conf. on Language Resources and Evaluation (LREC–08)*, Marrakech, Morocco, May 2008.
- G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, **A. Zlatintsi** and Y. Avrithis. **Movie Summarization Based on Audiovisual Saliency Detection**. *In Proc. Int'l Conf'. on Image Processing (ICIP–08)*, San Diego, CA, U.S.A., pages 2528-2531, Oct. 2008.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

More information please contact: *[nzlat,maragos]@cs.ntua.gr*

European Union European Social Fund — OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING investing in knowledge society — MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS MANAGING AUTHORITY — Co-financed by Greece and the European Union — NSRF 2007-2013 programme for development EUROPEAN SOCIAL FUND