EXTENDED PIPELINE FOR CONTENT-BASED FEATURE ENGINEERING IN MUSIC GENRE RECOGNITION

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Tina Raissi † Alessandro Tibo* Paolo Bientinesi †

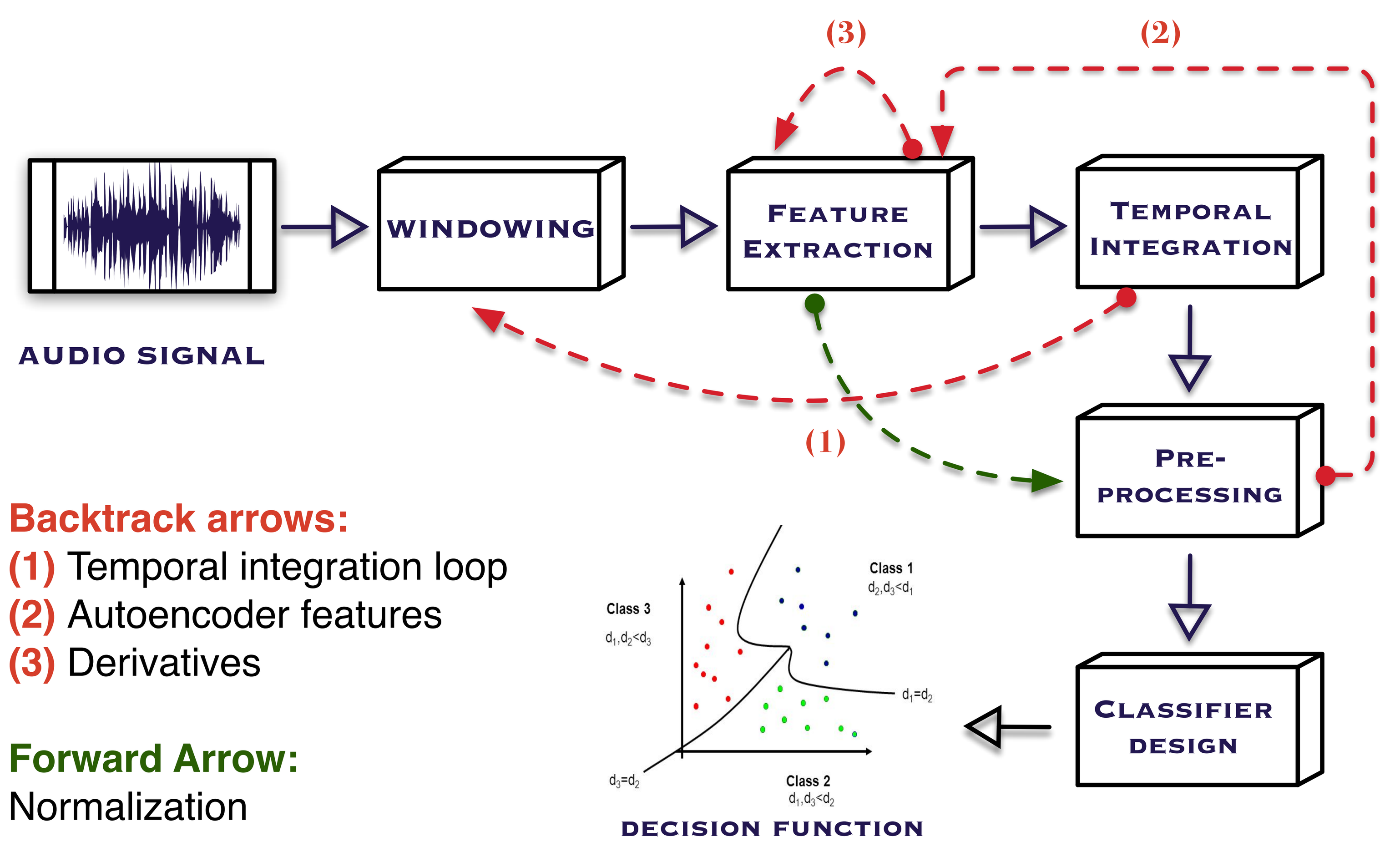
† RWTH Aachen University, Aachen Institute for Advanced Study in Computational Engineering Science, Germany
* University of Florence, Department of Information Engineering, Italy

RWTH AACHEN UNIVERSITY
HPAC High Performance and Automatic Computing

Research goal

- Development of a suitable representation of the intrinsic characteristics of the musical signal for automatic music genre recognition.
- Extension of the traditional two-step process of extraction and classification with additional independent stages.

Pipeline



Backtrack arrows:
(1) Temporal integration loop
(2) Autoencoder features
(3) Derivatives

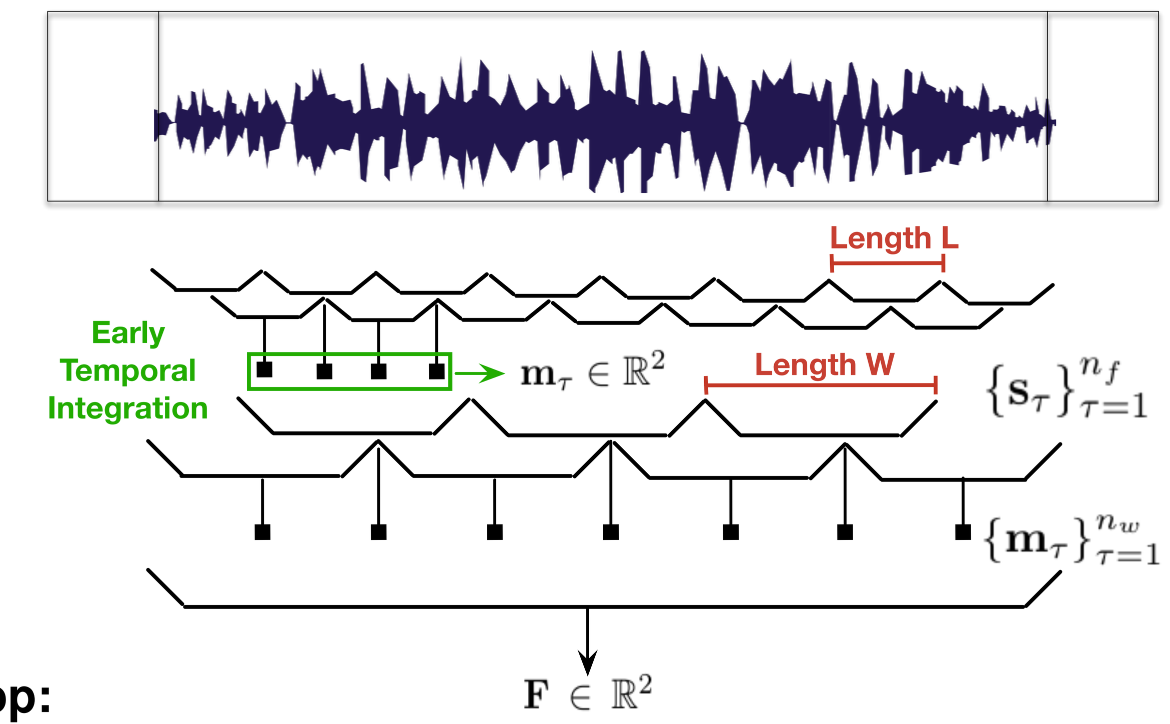
Forward Arrow:
Normalization

Content-based Feature Taxonomy

- CONTENT-BASED AUDIO FEATURES
  - PHYSICAL
    - TIME DOMAIN
      - ZERO CROSSING
      - ENERGY
      - ENTROPY OF ENERGY
      - ROOT MEAN SQUARE
      - FRACTION OF LOW ENERGY WINDOWS (FOLEW)
      - BEAT HISTOGRAM
    - FREQUENCY DOMAIN
      - SPECTRAL FLUX
      - SPECTRAL CENTROID
      - SPECTRAL ROLLOFF
      - SPECTRAL SPREAD
      - SPECTRAL VARIABILITY
      - LINEAR PREDICTION COEFFICIENTS
      - COMPACTNESS
      - CHROMOGRAM
  - PERCEPTUAL → CEPSTRAL DOMAIN → MEL FREQUENCY CEPSTRAL COEFFICIENTS

Feature Extraction, and Temporal Integration

Objective: Computation of $F \in \mathbb{R}^2$ for one feature vector, by performing the windowing, feature extraction and temporal integration steps twice.



First loop:

- **Windowing**: Signal split into analysis frames of $L = 50$ milliseconds.
- **Feature extraction**: Short-time feature values and their derivatives are extracted [1].
- **Temporal integration**: Application of MeanVar model [2], for medium-time vectors $\mathbf{m}_\tau = [\mu_\tau, \sigma_\tau]^T$.

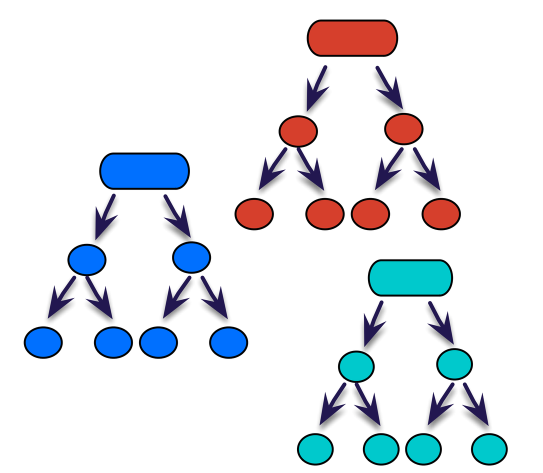
Second loop:

- **Windowing**: Texture window $W = 1$ second.
- **Feature extraction**: Derivation of FoLEW.
- **Temporal integration**: Vector $\mathbf{F}_k \in \mathbb{R}^2$, as average value of rows of $\mathbf{m}_\tau$.

Preprocessing: Feature Selection

Objective: Use the average variation of information entropy during the construction of a *Random Forests*, and select the best features.

- Denote by IG_{ij} the *information gain* of the i-th feature when splitting on j-th node,
- $IG_i = \sum_{j=1}^{|nodes|} IG_{ij}$ the IG of the feature, averaged over all trees,
- Eliminate the i-th feature if $IG_i = 0$.

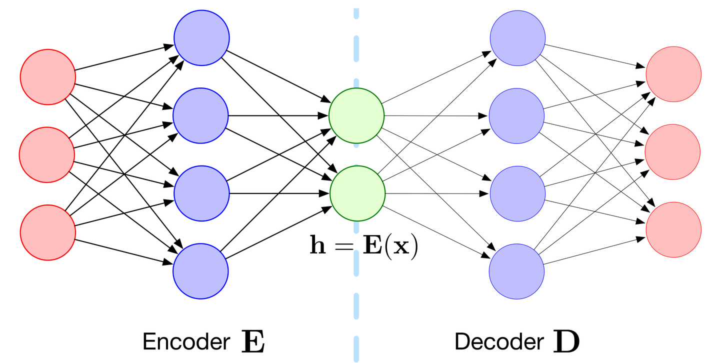


RANDOM FORESTS CLASSIFIER

Feature Extraction: Bottleneck Layer

Objective: Learn a low dimensional representation of the data, using a symmetric *autoencoder*

- Encoder $\mathbf{E}(\mathbf{x}) = \mathbf{h}$ with input $\mathbf{x} \in \mathbb{R}^n$,
- Decoder $\mathbf{D}(\mathbf{h}) = \mathbf{x}'$ with output $\mathbf{x}' \in \mathbb{R}^n$,
- Take the hidden embedding $\mathbf{h} \in \mathbb{R}^d$.



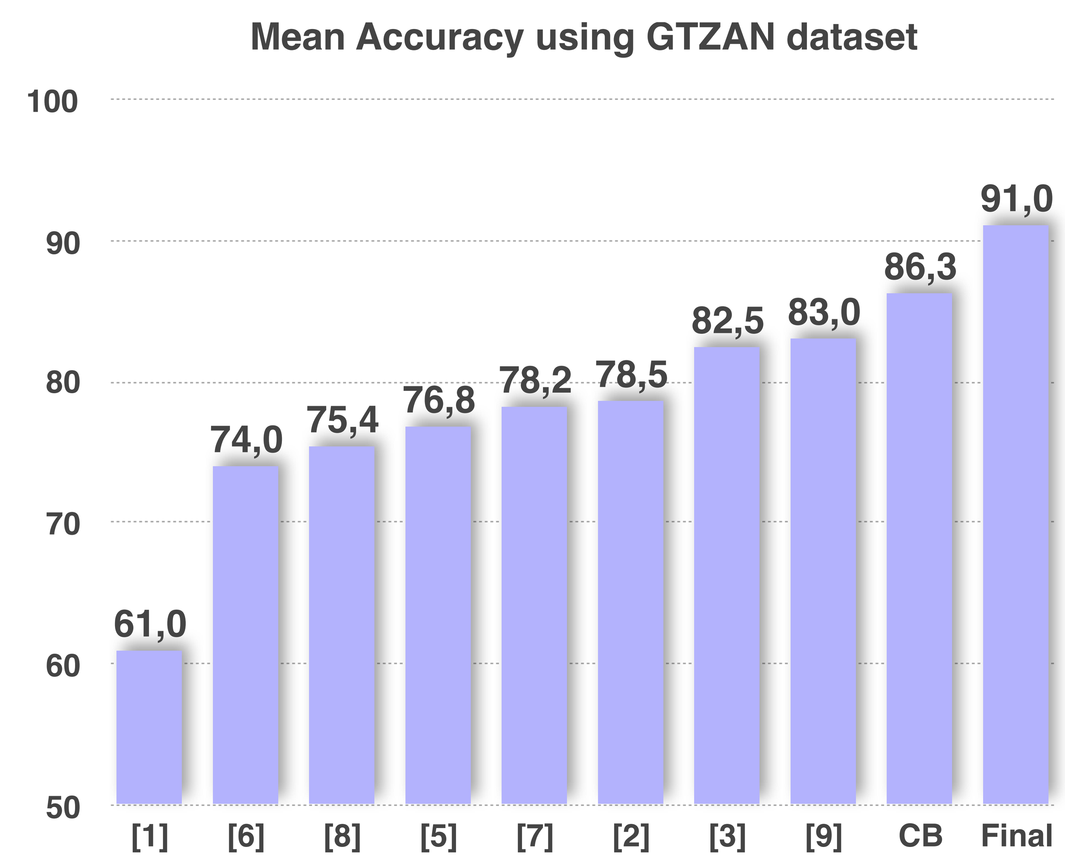
Binary cross-entropy loss:

$$\boxed{\mathcal{L}(\mathbf{x}, \mathbf{D}(\mathbf{E}(\mathbf{x}))) = -\big(\mathbf{x}\log(\mathbf{x}') + (\mathbf{1} - \mathbf{x})\log(\mathbf{1} - \mathbf{x}')\big)}$$

Classification

- The final dataset rescaled in $[0, 1]$.
- Support Vector Machines with radial basis kernel.
- Hyperparameter optimization with 10-fold cross-validation on training set.

Results



Our approach:

- **CB**: Content-based features, after feature selection step.
- **Final**: With addition of bottleneck's layer features.
- 10 random splits and 10-fold cross-validation.

Literature

[1] G. Tzanetakis, G. and P. Cook, IEEE Transactions on Audio, Speech, and Language Processing (2002)
[2] T. Li, M. Ogihara and Q. Li, ACM SIGIR (2003).
[3] J. Bergstra, N. Casagrande, D. Erhan, D. Eck and Balázs Kégl, Springer, J. Machine learning (2006).
[4] A. Meng, P. Ahrendt, J. Larsen, L. K. Hansen, IEEE Transactions on Audio, Speech, and Language Processing (2007).
[5] T. Lidy, A. Rauber, A., Pertusa and J. M. Inesta, Mirex (2007).
[6] A. Holzapfel and Y. Stylianou, IEEE Transactions on Audio, Speech, and Language Processing (2008).
[7] I. Panagakis, E. Benetos and C. Kotropoulos, ISMIR (2008).
[8] E. Benetos and C. Kotropoulos, IEEE Signal Processing Conference (2008).
[9] B.L. Strum, IEEE ICME, (2013).