Summary - Automatic music callsification and summarization

SVM Classifier for music with/without lyrics

Summarization : Task is to define a significant extract from a music : 5 min -> 20-30 seconds

Pure and vocal : classification with :
LPCC et LPC (Linear prediction methods) We can upgrade these methods by dividing the audio spectrum in 10 part an applying them on each parts

Zero crossing Rate, higher in the case of vocal music
Classifier : SVM with gaussian Kernel

To find the relevant features (which will be input in the SVM classifier) they used frames of 20 ms on which they calculated the Zero Crossing Rate (ZCR), and the LPCC/LPC.

Music pre-processed (remove silence, creation of the frames), then extraction of the features of each frames
Difference between pure and vocal music processing is the feature used.

PREPROCESSING :
-removing silence with the "short-time energy fuction" : if it is cotinuously lower than a threshold, then it it removed.
-feature selecion :
-pure : power-related = mel-frequency cepstrum coefficients, amplitude enveloppe, power spectrum
-vocal : voice-related = LPC derived cepstrum coefficients, Zero crossing Rate, spectrum flux, cepstrum flux

1) Mel-frequency Cepstral Coefs:
Useful to recognize structure of music signal, pitch and frequency. Calculated on frames of 20 ms.
They are coefficients calculated out of the power coefficients of the FFT, wich are then passed by
a triangular bandpass filter bank.
These are then represented on a re-normalized domain called the "Mel scale".

2) Spectrum Flux :
Variation value of the spectrum between two frames : Difference between the Fourrier coefs of a given frequency
from a frame to the next one.

3) Cepstrum flux :
Similar to spectrum flux, Difference between cepstrum of two successive frames.

4) Spectral power :
First, the frame's signal is weighted with a non-rectangular window function : the Hanning window.
Then, the spectral power is calculated with a formula, which give a max of 96dB.

5) Amplitude enveloppe :
Describes the energy change of the signal (function of the time), equivalent to the ADSR (=Attack, Decay,
Sustain, Release). Enveloppe is determined with a frame by frame RMS, the filtered with a Butterworth low-pass
filter of the third order and a cutoff frequency of 350, 1200 and 1700 Hz (empirical)

CLUSTERING AND elaboration of the summary :

1/ Pre-processing

Extract the features into a vector for each extract of 20 ms and *p%* of overlapping (p varies from 1% to 6%). Then, the distance between each vector is calculated into a distance matrix.

2/ Adaptative clustering with a target number of clusters.

- Begins with N clusters of 1 vector
- At each step it merges the two closest cluster into one cluster
- Note that we can use the barycenter of each cluster to minimize the noise.

At the end of the clustering algorithm, we got N clusters that reassembles the similar parts of the music. Doing so we know when variations occur in the music.

Evaluation methods are not relevant since they rely on genre recognition by humans to evaluate the quality of the summary.