COMPUTE AUDIO FEATURES
V1.0

Auteurs : Ewen Camberlein & Romain Lebarbenchon

TABLE DES MATIERES

Auteurs : Ewen Camberlein & Romain Lebarbenchon

1. Présentation du logiciel « compute_audio_features »

Ce logiciel permet d'extraire des descripteurs audio pour chaque canal d'un fichier au format « wav ».

Ce logiciel génère un fichier texte (« .txt ») pour chaque canal du fichier audio traité. Une ligne d'un de ces fichiers textes correspond à un vecteur de descripteur audio, calculé sur une trame du fichier. Pour chaque trame audio, ce vecteur composé de 78 descripteurs est calculé (MFCC avec leurs dérivées premières et secondes, dérivées première et seconde de l'énergie de la trame, centroïd spectral, étalement spectral, platitude spectrale et ZCR). Ces descripteurs sont encodés avec une précision de 15 digits, le séparateur utilisé étant la tabulation.

Les 78 descripteurs sont présentés dans l'ordre suivant (pour chaque ligne du fichier) :

- 24x MFCC
- 24x Δ MFCC
- 24x ΔΔ MFCC
- 1x Δ Energie de la trame
- 1x ΔΔ Energie de la trame
- 1x Centroïd spectral
- 1x Etalement spectral
- 1x Platitude spectrale
- 1x ZCR

2. Installation et utilisation du logiciel

L'exécutable à utiliser dépend de votre OS (cf. dossier correspondant). Ces exécutables ont été générés en utilisant le compilateur Matlab issu de la version Matlab R2015b. Pour les utiliser, il est donc nécessaire, au préalable, de télécharger et installer le MCR Matlab correspondant (à la fois à votre OS et à la version Matlab R2015b).

Les différents MCR sont disponibles à l'adresse suivante :
https://fr.mathworks.com/products/compiler/mcr/

Remarque : ces exécutables ne sont disponibles que pour des architectures 64 bits !

La syntaxe d'appel est la suivante sous windows :

```
> compute_audio_features.exe monFichierAudio.wav
```

Sous linux, il convient d'ajouter en complément le chemin vers le MCR de Matlab :

```
> ./run_compute_audio_features.sh CheminVersLeMcrMatlab monFichierAudio.wav
```

3. Glossaire

Acronyme	Définition
ASC	Audio Spectrum Centroïd
ASP	Audio Spectrum Spread
DCT	Discrete Cosine Transform
DSP	Densité Spectrale de Puissance
FFT	Fast Fourier Transform
MFCC	Mel Frequency Cepstral Coefficient
ZCR	Zero Crossing Rate

4. Fonctionnement détaillé

4.1. Description générale

La fonction d'extraction de descripteurs a pour but d'extraire du signal audio des paramètres permettant de le caractériser. Les descripteurs sélectionnés sont largement utilisés dans la littérature [1] [2] [3] [4]. L'implémentation réalisée suit celle préconisée par la norme MPEG-7 [5] quand cela est possible. Le signal audio est traité par trames de 32 ms avec un recouvrement de 50%.

Pour chaque trame de signal les descripteurs suivants sont extraits :

- MFCC
- Δ MFCC
- ΔΔ MFCC
- Δ Energie de la trame
- ΔΔ Energie de la trame
- Centroïd spectral (implémentation MPEG-7)
- Etalement spectral (implémentation MPEG-7)
- Platitude spectrale (implémentation dérivée de MPEG-7)
- ZCR

L'ajout d'informations temporelles est réalisé par le calcul de dérivées premières et secondes (communément appelées Δ et ΔΔ) sur les coefficents MFCC ainsi que sur la valeur d'énergie de chaque trame de signal.

Remarque : Avant tout traitement (sauf pour le calcul de MFCC qui nécessite une étape de « pré-emphasis » au préalable), une fenêtre de Hamming est appliquée sur chaque trame de signal afin de limiter l'influence des harmoniques parasites dans le domaine fréquentiel. Cette fenêtre est définie par :

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & pour\ n \in [0, N-1] \\ 0 & sinon \end{cases}$$

Avec N la taille de la fenêtre

Le signal fenêtré s'écrit de la manière suivante :

$$x_{win}(n) = x(n) * w(n)$$

Pour l'obtention du centroïd spectral, de l'étalement spectral et de la platitude spectrale un calcul de DSP est nécessaire sur la trame courante de signal. Celle-ci est calculée de la manière suivante :

$$DSP = |FFT(x_{win})|^2$$

Avec FFT l'opération de transformée de Fourier à court terme.

4.1.Calcul des descripteurs

4.1.1. Centroïd spectral

Le centroïd spectral représente le centre de gravité fréquentiel de la DSP d'un signal. En suivant les recommandations de la norme MPEG-7, sa formulation est la suivante :

$$\text{Asc} = \frac{\sum_{k=1}^{\frac{NFFT}{2}+1}[AscLogScale(k) * DSP(k)]}{\sum_{k=1}^{\frac{NFFT}{2}+1} DSP(k)}$$

En considérant :

$NFFT$: la taille de la FFT
$DSP(k)$: la puissance spectrale du k^{ième} bin fréquentiel

$AscLogScale(k)$ la fréquence du k^{ième} bin fréquentiel exprimée sur une échelle en octave :

$$AscLogScale(k) = log_2\left(\frac{f(k)}{1000}\right)$$

Avec $f(k)$ la fréquence du k^{ième} bin fréquentiel.

4.1.2. Etalement spectral

L'étalement spectral est une mesure de concentration de la puissance du signal autour du centroïd spectral. En suivant les recommandations de la norme MPEG-7, sa formulation est la suivante :

$$\text{Asp} = \sqrt{\frac{\sum_{k=1}^{\frac{NFFT}{2}+1}\left[\left(log_2\left(\frac{f(k)}{1000}\right) - Asc\right)^2 * DSP(k)\right]}{\sum_{k=1}^{\frac{NFFT}{2}+1} DSP(k)}}$$

4.1.3. Platitude spectrale

La platitude spectrale est une mesure de concentration énergétique de la représentation fréquentielle du signal. La mesure utilisée est dérivée de celle proposée par la norme MPEG-7. Au lieu de calculer une mesure de platitude spectrale pour chaque sous bande fréquentielle (précision non nécessaire pour notre application), nous calculons une seule valeur à partir des énergies moyennes de toutes les sous bandes fréquentielles.

Le calcul de ce descripteur s'effectue en 2 étapes :

- Calculer une énergie moyenne par sous bande :

  La norme MPEG-7 préconise d'utiliser des sous bandes d'un quart d'octave avec un recouvrement partiel entre sous bandes successives.

  Calcul du nombre de sous bandes (à arrondir à l'entier inférieur) :

$$numBands = \frac{1}{AsfRes} * log_2\left(\frac{F_{min}}{(Fs/2)}\right)$$

Avec :

AsfRes = 0.25 la résolution en quart d'octave ;
F_{min} =250 Hz la fréquence minimum considérée, comme préconisé dans la norme ;
Fs : la fréquence d'échantillonnage.

Les indices des bins fréquentiels minimum et maximum pour la k[ième] sous bande s'expriment alors de la manière suivante (en arrondissant à l'entier le plus proche):

$$i_{min_k} = \left(F_{min} * 2^{\frac{(k-1)}{4}} * 0.95\right) * \frac{NFFT}{Fs} + 1$$

$$i_{max_k} = \left(F_{min} * 2^{\frac{k}{4}} * 1.05\right) * \frac{NFFT}{Fs} + 1$$

Pour chaque sous bande on calcule alors la valeur moyenne Esb_k des coefficients de la DSP compris entre les indices i_{min_k}et i_{max_k}.

- Calculer la mesure de platitude spectrale :

  La mesure de platitude spectrale s'exprime comme le rapport entre la moyenne géométrique et la moyenne arithmétique des énergies moyennes par sous bande.

$$Asp = \frac{\exp\left(\frac{1}{numBands}\sum_{k=1}^{numBands}\ln(Esb_k)\right)}{\frac{1}{numBands}\sum_{k=1}^{numBands}Esb_k}$$

4.1.4. ZCR (Zero Crossing Rate)

Le ZCR est le taux de passage par zéro. Il est calculé en comptant le nombre d'inversion de signe sur la durée de la trame de signal et en divisant cette quantité par la taille de la trame.

4.1.5. MFCC (Mel Frequency Cepstral Coefficients)

Les MFCC sont des descripteurs communément utilisés dans la littérature par exemple pour la reconnaissance de locuteur. Les MFCC sont des paramètres dont le calcul est basé sur une transformée en cosinus discrète (DCT) appliquée au spectre de puissance d'un signal, calculé sur des bandes de fréquences suivant une échelle de Mel. La construction de cette échelle est basée sur la perception de l'oreille humaine.

Le calcul des MFCCs est réalisé par les étapes suivantes :

- Pre-emphasis

  Cette étape permet d'augmenter l'énergie du signal dans les hautes fréquences. L'échantillon filtré $y(n)$ est obtenu à partir de l'échantillon courant $x(n)$ et du précédent $x(n-1)$tel que:

$$y(n) = x(n) - 0.95\, x(n-1)$$

- Fenêtrage

Ce signal filtré est ensuite découpé en trames de 32 ms avec un recouvrement de 50% entre trames. Chaque trame est ensuite fenêtrée par une fenêtre de Hamming comme décrit en 4.1.

- Calcul du spectre d'amplitude

  Le spectre d'amplitude est ensuite calculé en prenant le module de la FFT du signal fenêtré.

$$SA = |FFT(y_{win})|$$

- Découpage en sous bandes

  Ce découpage est effectué par un banc de M=32 filtres triangulaires sur une échelle de Mel (exemple ci-dessous pour un banc de filtre triangulaire à 10 sous bandes):

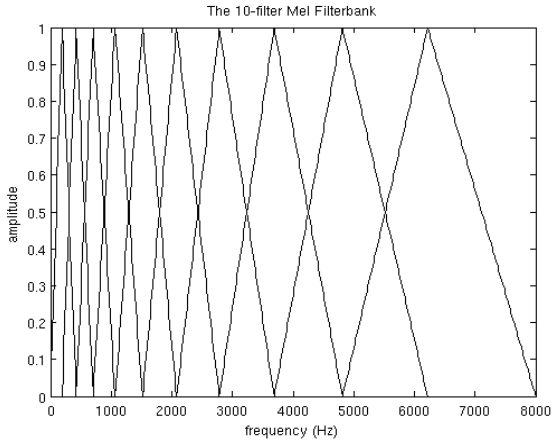

Figure 1 : Banc de filtres sur l'échelle de Mel

  En sommant les coefficients du spectre d'amplitude pondérés par les valeurs de chaque filtre triangulaire on obtient les valeurs des énergies en sous bande E_m.

- Discrete Cosine Transform (DCT)

  La dernière étape de l'algorithme est de calculer une DCT sur ces énergies en sous-bandes. Les coefficients ainsi obtenus sont les MFCCs et seuls les 24 premiers sont conservés.

$$X_k = \sum_{m=0}^{M-1} x_m \cos\left(\frac{\pi}{M}(m+\frac{1}{2})k\right)$$

  En considérant :

$$k \in \{1,...,24\}$$

$$x_m = \ln(E_m)$$

  Remarque :
  Le coefficient X_0 (k=0) n'est pas utilisé comme descripteur, afin d'éviter l'introduction d'une dépendance au « gain d'entrée », la valeur de ce coefficient étant liée à l'énergie de la trame courante.

Le flow chart suivant résume l'implémentation réalisée de l'algorithme de calcul des MFCC :

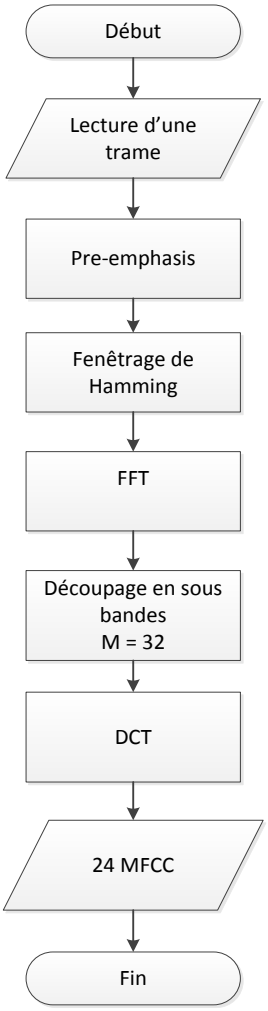

Figure 2 : Flow chart du calcul des MFCC

En plus des MFCC, les Δ et ΔΔ (dérivées premières et secondes) de ces coefficients sont également calculés et utilisés comme descripteurs (cf. description en 4.1.7)

4.1.6. Evolution de l'énergie

Sur chaque trame de signal $\{x_{win}(n)\}_{n\in[1,N]}$ (fenêtrée par la fenêtre de Hamming décrite précédemment), on réalise une mesure d'énergie de la manière suivante :

$$E_n = max\left(-100,10 * log_{10}\left[\frac{1}{N}\sum_{n=1}^{N}[x_{win}(n)]^2\right]\right)$$

Cette mesure n'est pas utilisée directement en tant que descripteur car elle dépend du gain d'entrée autant que du type de signal.

Les descripteurs retenus dans ce cadre sont les Δ et ΔΔ (dérivées premières et secondes) de cette mesure. Le calcul de ces derniers est expliqué dans le paragraphe suivant (4.1.7).

4.1.7. Evolution temporelle de descripteurs

Afin d'obtenir des informations sur l'évolution temporelle des descripteurs, on peut calculer leur dérivée première et seconde, communément appelées Δ et ΔΔ.

Pour calculer les coefficients deltas, la formule suivante est utilisée (approximation par un développement limité d'ordre N) :

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Avec d_t le coefficient Δ pour la trame t.
Avec c_t un descripteur à l'indice t.

En pratique, on utilise ici un développement limité d'ordre 2 :

$$d_t = \frac{-2c_{t-2} - c_{t-1} + c_{t+1} + 2c_{t+2}}{10}$$

Les coefficients ΔΔ (accélération) sont introduits de la même manière en utilisant les d_t nouvellement calculés à la place des c_t dans la formule ci-dessus.

5. Bibliographie

[1] D. A. Reynolds, T. F. Quatieri et R. B. Dunn, «Speaker Verification Using Adapted Gaussian Mixture Models,» *Digital Signal Processing 19-41*, n° %110, 2000.

[2] S. Ntalampiras, I. Potamitis et N. Fakotakis, «Automatic recognition of urban environmental sounds events.».

[3] G. Peeters, «A large set of audio features for sound description (similarity and classification) in the CUIDADO project,» IRCAM, Analysis/Sunthesis Team, 2004.

[4] J.-L. Rouas, J. Louradour et S. Ambellouis, «Audio events detection in public transport vehicule,» chez *IEEE Intelligent Transportation Systems Conference*, 2006.

[5] ISO/IEC JTC, MPEG-7 (Multimedia content description interface), 2009-10-30.