Projet Multimedia:
Détection des différentes phase de l'émission

Timothee, El Mostafa, Hamdi et Othman

October 2018

1 Résumé du cours

1.1 Analyse audio

L'analyse audio peut être réalisée sur trois type de contenu : parole, music ou bruit environnemental. De plus, dans ce domaine, les analyses sont très différentes lorsqu'il s'agit de parole ou de musique/bruit environnemental.
Il existe de nombreuses applications : classification de genre musicaux, analyse rythmique, audio encodage, reconnaissance vocale, linguistique... Dans le domaine de la parole, l'analyse audio représente entre autres une passerelle entre l'Homme et la Machine.
La plupart des applications repose sur des algorithmes de classification supervisé ou non supervisé.
Nous allons donc par la suite décrire les principales étapes d'une classification audio :

a. Digitization
b. Preprocession
c. Feature computation
d. Temporal integration
e. Classifier training

1.1.1 Digitization

Les analyses audios sont réalisées sur des signaux numériques qui été initialement des signaux acoustiques. Un signal acoustique est enregistré à partir d'un micro sou forme de signal analogique puis par une méthode d'échantillonnage nous obtenons un signal numérique. La quantité d'information est en générale codée sur 16 bits à une fréquence d'au moins deux fois 20 kHz pour ne pas perdre d'informations. En effet, la plage de fréquence audible pour un humain

est de 20 Hz à 20 kHz. On remarquera que le débit obtenu est de 16 bit x 40 k Hz = 640 bit / s. D'où l'utilité d'algorithme de compression (mp3, mp4)

1.1.2 Preprocessing

L'objectif du preprocessing est à la fois de filtrer les bruits sur le signal, réduire le taux d'information (eg. subsampling) et la normalisation.

1.1.3 Feature computation

Pour définir un vecteur temporel de features, il est important de bien définir la fenêtre d'analyse. La segmentation temporelle peut être soit statique à intervalle de temps constant (eg. 20 ms) ou dynamique pour varier en fonction de la quantité d'information. Il existe trois catégories de features : 1.

a. Temporal feature : description de la forme de l'onde du signal (eg. Zero Crossing Rate),

b. Spectral features : représentation fréquentielle du signal (eg. DFT, STFT, Cepstrum ou MFC),

c. Preceptual features : représentation basée sur des considérations psychoacoustiques.

Nous allons détailler ici les spectral features les plus utilisés en analyse audio :

- Le Short-Time Fourier Transform (STFT) est un outil puissant pour le processing du signal audio. Il permet de définir une classe particulièrement utile de distributions fréquence-temps qui spécifie des amplitudes complexe par rapport au temps et à la fréquence pour n'importe quel signal :
  - But : Extraire des informations d'un signal analogique, ou numérique
  - Choix de la fenêtre : La largeur du lobe principal permet de contrôler la précision fréquentielle. Plus il est étroit, meilleur est la résolution fréquentielle. Les lobes secondaires doivent pouvoir être négligés par rapport au lobe principal.
  - Choix de la taille de la fenêtre : Obligation de prendre une résolution fréquentielle en fonction de la taille de la fenêtre qui permette de s'assurer les composantes du signal étudié soient séparées par une distance suffisamment importante.
  - Limitations de la représentation spectrale: l'analyse est globale et ne permet pas de capturer l'information temporelle comme le début et la fin du signal ou l'apparition d'une singularité, l'analyse STFT exige de connaitre l'intégralité du signal. La représentation engendrée est de grande dimensionnalité.

- Le modèle Source-Filter modélise la parole comme une combinaison de sources indépendantes : une source de son par exemple le souffle pulmonaire qui fait entrer en vibrations les cordes vocales, et un filtre acoustique linéaire comme par exemple le canal vocal qui joue le rôle d'amplificateur de certaines fréquences.

  Les deux font des contributions séparées aux features caractéristiques du son résultant. La source dans le son de la voix dans le discours exprimé est responsable de la hauteur du son alors que le filtre à cordes vocales est responsable de l'emplacement des formants et de la forme spectrale excessive. Dans un vrai spectre de discours, la forme globale du filtre et l'emplacement des formants sont souvent noyés par les effets du spectre source. S'il était possible de supprimer les effets de source, les deux spectres pourraient être étudiés séparément pour donner une image plus précise des features du discours.
  Enlever les effets de la source du spectre a pour conséquence d'enlever les sources de variabilité du signal. Il y a plusieurs techniques pour séparer la source et le filtre dans un signal audio comme par exemple l'analyse Cepstrale.

  L'analyse Cepstrale repose sur l'observation qu'un spectre de discours logarithmique est constitué de la source et du spectre de filtres ajoutés ensemble. L'origine de cette idée est que le filtrage dans le domaine des fréquences est réalisé en multipliant les spectres ensemble.
  La multiplication correspondant à l'ajout de logarithmes et un spectre filtré peut être dérivé en ajoutant des spectres logarithmiques (ou dB). La procédure pour l'analyse Cepstrale est de prendre la transformation inverse de Fourier du spectre dB, convertissant ainsi le signal en un domaine temporel. Ce signal de domaine temporel n'est pas un signal acoustique régulier, puisqu'il a été dérivé du spectre logarithmique; pour cette raison, il est appelé une cepstrum. Puisque le spectre des dB est la somme de deux spectres, La propriété importante du cepstrum est qu'il est la somme de deux composants correspondants à la source et au filtre. La convolution permet donc de caractériser la transformation entrée/sortie réalisée par un filtre linéaire invariant. Le tout est un système invariant dans le temps. La partie inférieure du cepstrum correspond au filtre tandis que la partie supérieure (ou plutôt le milieu du cepstrum reflétée) correspond à la source

- L'analyse Mel fréquence du discours est basée les expériences de perception humaine. Il a été observé que l'oreille humaine agissait comme un filtre en se concentrant que sur certaines composantes de la fréquence. Ces filtres sont non uniformément espacés dans l'axe de fréquence : Plus de filtres dans les régions basses fréquences et moins de nombre de filtres en régions de haute fréquence. Les MFCC sont utilisés dans deux domaines :

  – La synthèse de discours :

        * Utilisé pour joindre deux segments de discours S1 et S2
        * Représentation de S1 comme une séquence de MFCC
        * Représentation de S2 comme une séquence de MFCC
        * ointure au point où les MFCC de S1 et S2 ont la distance euclidienne minimale
    - La reconnaissance vocale:
        * Les MFCC sont surtout utilisées comme features dans les systèmes de pointe de reconnaissance de discours

Pour résumer, le discours est analysé sur des courtes fenêtres d'analyse dans lesquelles le spectre est obtenu en utilisant la FFT. Le spectre passe après par un filtre Mel pour obtenir le spectre Mel (MelSpectrum), l'analyse Cepstrale est ainsi réalisée sur le spectre Mel pour obtenir les coefficients Mel-Frequence Cepstrales. En représentant le discours comme une séquence de vecteurs Cepstraux, on peut donner ces vecteurs au classifieurs pour réaliser la reconnaissance vocale.

Pour savoir quels features utiliser dans le cadre d'une analyse précise, nous pouvons nous baser sur un jugement d'expert, un algorithme de sélection automatique (Forward and Backward Stepwise selection) ou une méthode de réduction de la dimensionnalité (Wrapper, Filter or Embedded methods). Il existe aussi les réseaux de neurones qui, en se basant sur des raw data, permettent d'éviter ce type de questionnement. Temporal integration Une fois le vecteur temporel de features défini, nous effectuons une intégration temporelle (eg. La moyenne sur 30 s). L'intégration permet de :

- Supprimer les bruits et d'améliorer la robustesse du modèle,
- Synchroniser les features avec le bon choix de fenêtre,
- Capturer l'évolution temporelle des features.

Classifier training Nous définissons alors un échantillon d'apprentissage pour entrainer notre classifier et de test pour valider la justesse de la calibration du classifier en évitant l'overfitting