M.C. Doing Studio
Mathematical Department
Sun Yat-sen University

Acoustic Recognition Summary

Version 0.0.4

February 4, 2017

Updates

Version	**Authors**	**Date**	**Remark**
0.0.1	Zheng Rockman	January 30, 2017	First draft
0.0.2	Zheng Rockman	January 31, 2017	Add notations and concepts
0.0.3	Zheng Rockman	February 2, 2017	Add notations and concepts
0.0.4	Zheng Rockman	February 4, 2017	Supplement MFCC

Contents

1 Introduction

2 Notation and Concepts

- $x_a(t)$, an **analog signal** as a function varying continuously in time, where subscript a stands for analog.
- $x[n] = x_a(nT)$, a **discrete-time signal**, if we sample the signal $x_a(t)$ with a sampling period T. **Sampling** means evaluating or measuring the original analog signal at discrete points.
- $F_s = 1/T$, **sampling frequency**, where subscript s stands for sampling.
- A **digital system** T is a transform that, given an input signal $x[n]$, generates an output signal $y[n]$:

  $$y[n] = T\{x[n]\}$$

  .
- A digital system T is defined to be **linear** iff

  $$T\{ax[n] + by[n]\} = aT\{x[n]\} + bT\{y[n]\},$$

  for any values of a, b and any signals $x[n], y[n]$.
- A digital system is **time-invariant** if

  $$y[n - n_0] = T\{x[n - n_0]\}.$$

- A digital system which is **linear time-invariant (LTI)** can be uniquely described by:

  $$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n] * h[n],$$

  where $*$ is defined as the **convolution** operator.
- LTI systems are completely characterized by the signal $h[n]$, which is known as the system's **impulse response** because it is the output of the system when the input is an **impulse** $\delta[n]$.
- Some useful digital signals. **Kronecker delta** or **unit impulse**:

  $$\delta[n] = \begin{cases} 1, & n = 0; \\ 0, & \text{otherwise.} \end{cases}$$

  Unit step:

  $$u[n] = \begin{cases} 1, & n \geq 0; \\ 0, & n < 0. \end{cases}$$

- **discrete-time Z transform** of signal $x[n]$:

$$X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n}$$

- **discrete-time Fourier transform** of signal $x[n]$:

$$X\left(\mathrm{e}^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x[n] \mathrm{e}^{-j\omega n}.$$

- The **inverse discrete-time Fourier transform** is defined as

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X\left(\mathrm{e}^{j\omega}\right) \mathrm{e}^{j\omega n} \mathrm{d}\omega.$$

  The Fourier transform is invertible.

- The **real cepstrum** of a digital signal $x[n]$ is defined as:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left| X\left(\mathrm{e}^{j\omega}\right) \right| \mathrm{e}^{j\omega n} \mathrm{d}\omega.$$

- In signal processing, a **filter** is a device or process that removes some (often unwanted) components or features from a signal. In particular, filter often refers to function that eliminates some parts of a signal when they interact with each other (by multiplication or convolution). Filters that allow low frequency components of a signal to pass but block out its high frequency parts are called **low-pass**. In contrast, filters that pass high frequency parts of a signal but discard low frequency components are called **high-pass**. Besides, **band-pass** filters only allow a particular band of a signal to pass.

- The Fourier transform $H\left(\mathrm{e}^{j\omega}\right)$ of a filter $h[n]$ is called the system's **frequency response** or **transfer function**.

- It is useful to find an impulse response $h[n]$ whose Fourier transform is

$$H\left(\mathrm{e}^{j\omega}\right) = \begin{cases} 1, & |\omega| < \omega_0; \\ 0, & |\omega| \geq \omega_0. \end{cases}$$

  This $H\left(\mathrm{e}^{j\omega}\right)$ is the **ideal low-pass filter** because when we multiply it with a signal in frequency domain, it lets all frequencies below ω_0 pass through unaffected and completely blocks frequencies above ω_0. It is called ideal because of its simple form. Using the definition of Fourier transform, we obtain:

$$h[n] = \frac{\omega_0}{\pi} \mathrm{sinc}\left(\omega_0 n\right),$$

  where

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}.$$

- **Window functions** are signals that are concentrated in time, often of limited duration. Window functions are also concentrated in low frequencies.

- The **rectangular window** is defined as

$$h_\pi[n] = u[n] - u[n-N].$$

- The **generalized Hamming window** is defined as:

$$h_h[n] = \begin{cases} (1-\alpha) - \alpha\cos(2\pi n/N), & 0 \geq n < N; \\ 0, & \text{otherwise.} \end{cases}$$

  or can be expressed in terms of the rectangular window as:

$$h_h[n] = h_\pi[n]\left[(1-\alpha) - \alpha\cos(2\pi n/N)\right].$$

  When $\alpha = 0.5$ the window is known as the **Hanning window**, whereas for $= 0.46$ it is the **Hamming window**.

- A **filterbank** is a collection of filters that span the whole frequency spectrum.

- A new set of techniques called **short-time analysis** or **short-time Fourier analysis** are proposed to compute spectrogram from its corresponding time signal. These techniques decompose the speech signal into a series of short segments, referred to as **analysis frames**, or simply **frames**, and analyze each one independently.

- Given a signal $x[n]$, we define the short-time signal $x_m[n]$ of the m-th frame as

$$x_m[n] = x[n]w_m[n],$$

  the product of $x[n]$ by a window function $w_m[n]$, which is zero everywhere except in a small region corresponding to that frame. While the window function can have different values for different frames m, a popular choice is to keep it constant for all frames:

$$w_m[n] = w[m-n],$$

  where $w[n] = 0$ for $|n| > N/2$.

- With the above framework, the short-time Fourier representation for frame m is defined as

$$X_m\left(\mathrm{e}^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x_m[n]\mathrm{e}^{j\omega n} = \sum_{n=-\infty}^{\infty} w[m-n]x[n]\mathrm{e}^{j\omega n},$$

  with all the properties of Fourier transforms

- A **spectrogram** of a time signal is a special two-dimensional representation that displays time t in its horizontal axis and frequency f in its vertical axis. A gray scale is typically used to indicate the energy at each point (t, f) in spectrogram with white representing low energy and black high energy.

- The **Mel-Frequency Cepstrum Coefficients (MFCC)** is a representation defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal. Given the DFT of the input signal

$$X_a[k] = \sum_{n=0}^{N-1} x[n]\mathrm{e}^{-j2\pi nk/N}, 0 \leq k < N,$$

we define a filterbank with M filters ($m = 1, 2, , M$), where the m-th filter is a triangular filter given by:

$$H_m[k] = \begin{cases} 0, & k < f[m-1]; \\ \frac{k-f[m-1]}{f[m]-f[m-1]}, & f[m-1] \leq k \leq f[m]; \\ \frac{f[m+1]-k}{f[m+1]-f[m]}, & f[m] \leq k \leq f[m+1]; \\ 0, & k > f[m+1]. \end{cases}$$

which satisfies $\sum_{m=0}^{M-1} H_m[k] = 1$. Define f_l and f_h to be the lowest and highest frequencies of the filterbank in Hz, F_s the sampling frequency in Hz, M the number of filters, and N the size of the FFT (the number of samples). The boundary points $f[m]$ are uniformly spaced in the mel-scale:

$$f[m] = \frac{N}{F_s} B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right),$$

where the mel-scale B is given by:

$$B(f) = 1125 \ln(1 + f/700),$$

and B^{-1} is its inverse

$$B^{-1}(b) = 700(\exp(b/1125) - 1).$$

We then compute the log-energy at the output of each filter as:

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], 0 \leq m < M.$$

The mel frequency cepstrum coefficients (MFCC) are then the discrete cosine transform of the M filter outputs:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\frac{\pi n(m+1/2)}{M}\right), 0 \leq n < M,$$

where M varies for different implementations from 24 to 40. For speech recognition, typically only the first 13 cepstrum coefficients are used.