Speech Recognition and Speech Translation

Digital Feature Representations for Automatic Speech Recognition
Summer 2018

Stefan Riezler

Computational Linguistics & IWR
Heidelberg University, Germany
riezler@cl.uni-heidelberg.de

Fourier Analysis: Mathematics of Spectral Analysis

- Any complex function can be described as **summation of sinusoidal functions of increasing frequency** (Jean-Baptiste Joseph Fourier, 1822)
- Fourier transformation is **transformation of time varying signals into frequency space**
- Based on complex numbers since complete description gives magnitude at frequency band (real part) and phase angle (imaginary part)

Complex Numbers

- Complex number in Cartesian coordinate system: $R + iI$
  - Consists of real part R and imaginary part I, where $R, I \in \mathbb{R}$, $i = \sqrt{-1}$
  - Numbers sitting on plane above/below real numbers ($I = 0$)

Complex Numbers

- Complex number in Cartesian coordinate system: $R + iI$
  - Consists of real part R and imaginary part I, where $R, I \in \mathbb{R}$, $i = \sqrt{-1}$
  - Numbers sitting on plane above/below real numbers ($I = 0$)
- Polar coordinate representation, determined by radius r and angle θ, where $r = \sqrt{R^2 + I^2}$, $\theta = \arctan(\frac{I}{R})$

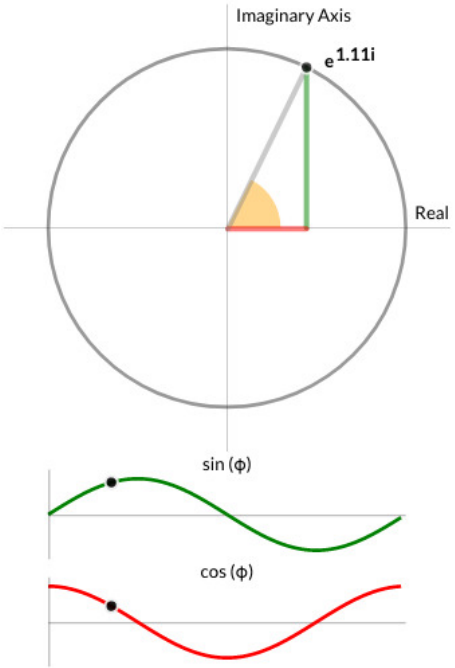
Complex Numbers

- Complex number in Cartesian coordinate system: $R + iI$
  - Consists of real part R and imaginary part I, where $R, I \in \mathbb{R}$, $i = \sqrt{-1}$
  - Numbers sitting on plane above/below real numbers ($I = 0$)
- Polar coordinate representation, determined by radius r and angle θ, where $r = \sqrt{R^2 + I^2}$, $\theta = \arctan(\frac{I}{R})$
  - Equivalence: $r(\cos(\theta) + i\sin(\theta)) = r(\frac{R}{r} + i\frac{I}{r}) = R + iI$

Complex Numbers

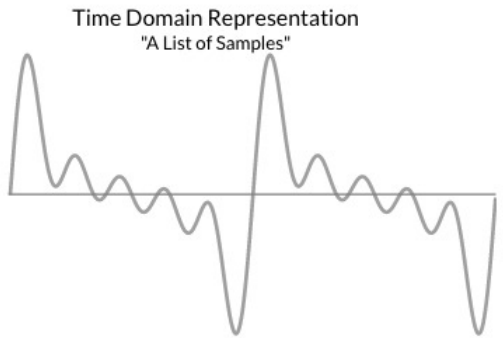
- Complex number in Cartesian coordinate system: $R + iI$
  - Consists of real part R and imaginary part I, where $R, I \in \mathbb{R}$, $i = \sqrt{-1}$
  - Numbers sitting on plane above/below real numbers ($I = 0$)
- Polar coordinate representation, determined by radius r and angle θ, where $r = \sqrt{R^2 + I^2}$, $\theta = \arctan(\frac{I}{R})$
  - Equivalence: $r(\cos(\theta) + i\sin(\theta)) = r(\frac{R}{r} + i\frac{I}{r}) = R + iI$
- Euler's formula:
  - $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, for all $\theta \in \mathbb{R}$
  - Complex number: $re^{i\theta}$
  - r is magnitude, θ is phase angle

Complex Phasors



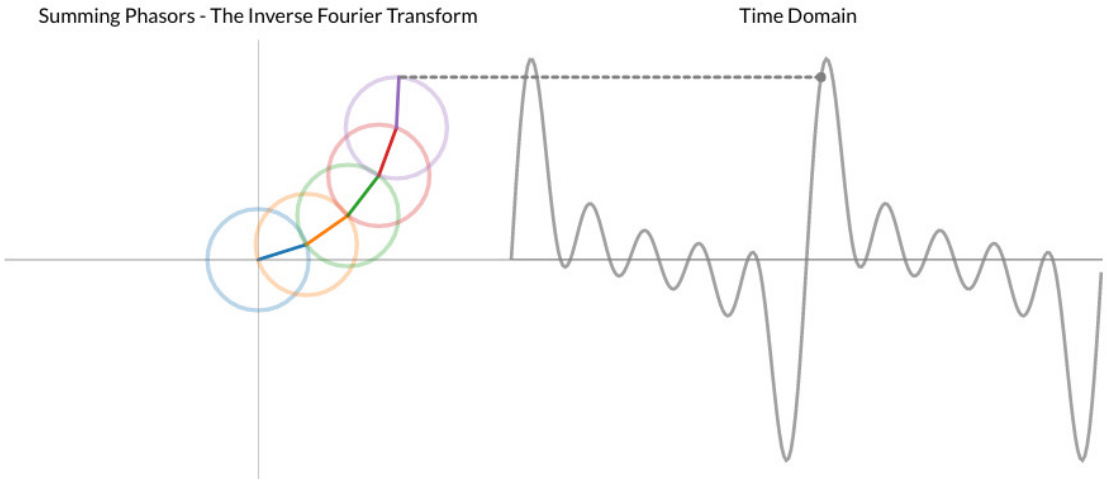
- Complex numbers visualized as **complex phasors in frequency domain**, corresponding to **sinusoidal waves in time domain**
- real axis given by cosine, imaginary axis given by sine

Fourier Transformation into Sum of Phasors



- Components in frequency domain visualized as list of phasors
  - Frequency visualized by speed of rotation, amplitude by size of radius
  - Phasors are harmonically related, i.e., frequencies are multiples of each other
- Goal: Reproduce time domain signal by summing phasors

Fourier Transformation into Sum of Phasors



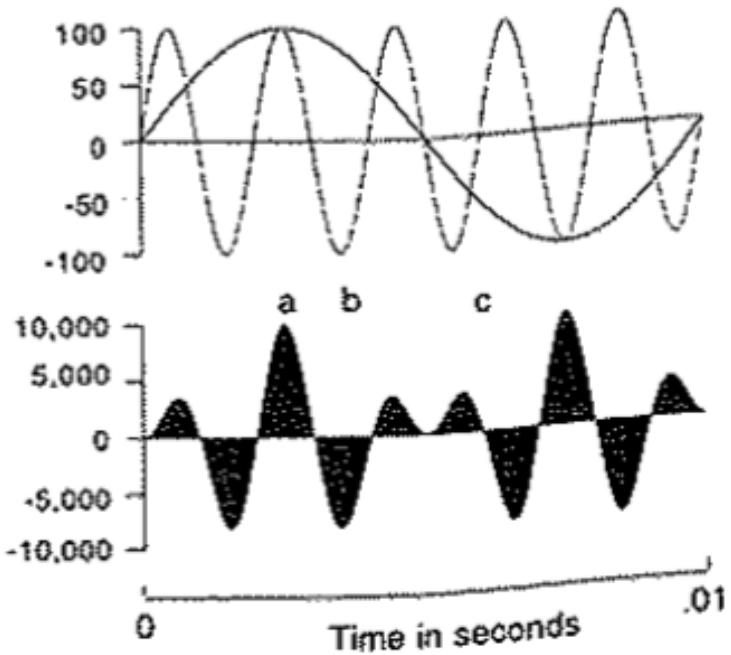
- Summation is visualized by attaching center of next phasor at tip of previous phasor
- Time domain is reproduced by vertical distance of tip of last phasor to origin
- Mathematically: $a_1 \sin(f_1 * \theta) + a_2 \sin(f_2 * \theta) \ldots$ where a_i and f_i are amplitude and frequency of i-th sine

Discrete Fourier Transform (DFT)

$$DFT(k) = \sum_{n=0}^{N-1} x[n] e^{-i\theta}$$
$$= \sum_{n=0}^{N-1} x[n](\cos(\theta) - i\sin(\theta)), \text{ where } \theta = \frac{2\pi kn}{N}$$

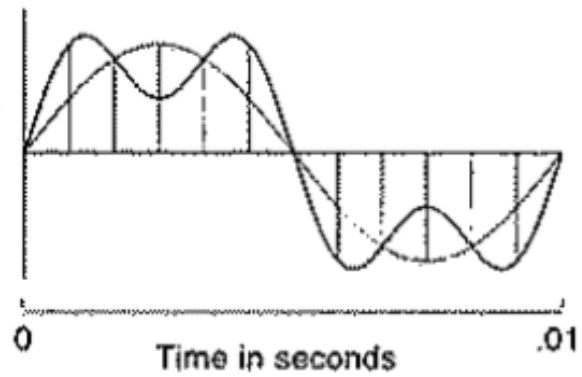
- $x[n]$ is signal measured at time n out of N samples per cycle
- $DFT(k)$ transforms sequence $x[n]$ into magnitude and phase of discrete frequency component k for $k = 0, \ldots, N-1$
- At the core: Measure **correlation** of complex wave with sinusoidal components, by taking **dot products** of input signal with sinusoidal waves of varying frequency

Dot Products



- Sinusoidal waves with different frequencies (here: 100 Hz and 500 Hz) are orthogonal, i.e., zero correlation
- The only correlation of a sinusoidal wave and a complex wave will be with components at same frequency

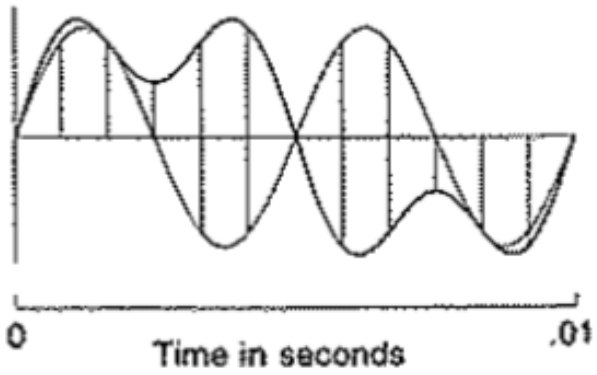
Dot Products



Time(s)	Complex Amplitude	100 Hz Amplitude	Product	Accumulating Sum of Products
0.00083	60	30	1800	1800
0.00167	52	52	2704	4504
0.00250	30	60	1800	6304
0.00333	52	52	2704	9008
0.00417	60	30	1800	10808
0.00500	0	0	0	10808
0.00583	−60	−30	1800	12608
0.00667	−52	−52	2704	15312
0.00750	−30	−60	1800	17112
0.00833	−52	−52	2704	19816
0.00917	−60	−30	1800	21616
0.01000	0	0	0	21616

- Complex wave at fundamental frequency 100Hz, sine at 100Hz

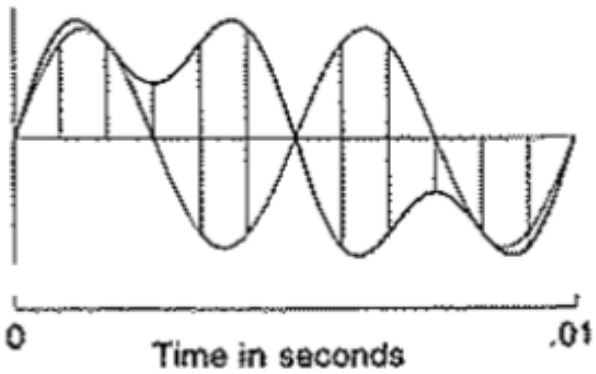
Dot Products



Time(s)	Complex Amplitude	200 Hz Amplitude
0.00083	60	52
0.00167	52	52
0.00250	30	0
0.00333	52	−52
0.00417	60	−52
0.00500	0	0
0.00583	−60	52
0.00667	−52	52
0.00750	−30	0
0.00833	−52	−52
0.00917	−60	−52
0.01000	0	0

- Complex wave at fundamental frequency 100Hz, sine at 200Hz

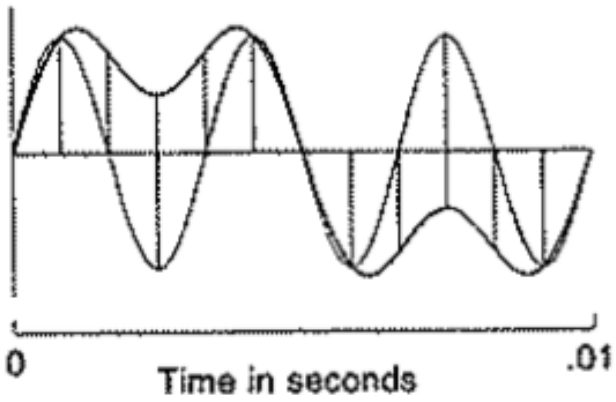
Dot Products



Time(s)	Complex Amplitude	200 Hz Amplitude	Product	Accumulating Sum of Products
0.00083	60	52	3120	3120
0.00167	52	52	2704	5824
0.00250	30	0	0	5824
0.00333	52	−52	−2704	3120
0.00417	60	−52	−3120	0
0.00500	0	0	0	0
0.00583	−60	52	−3120	−3120
0.00667	−52	52	−2704	−5824
0.00750	−30	0	0	−5824
0.00833	−52	−52	2704	−3120
0.00917	−60	−52	3120	0
0.01000	0	0	0	0

- Complex wave at fundamental frequency 100Hz, sine at 200Hz

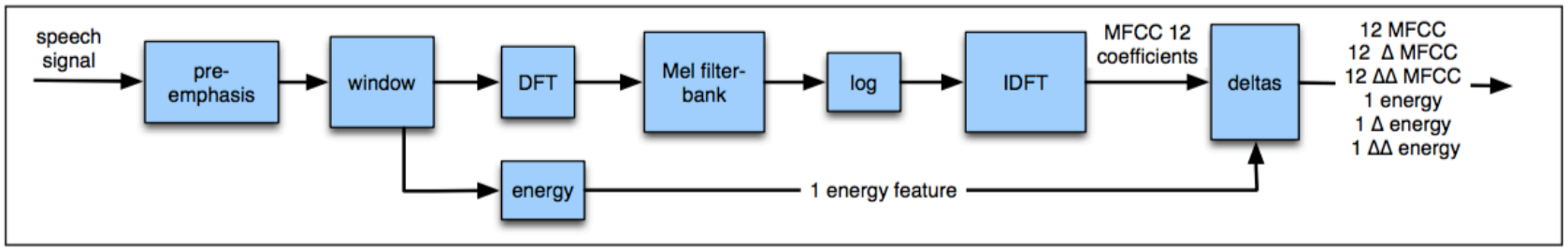
Dot Products



Time(s)	Complex Amplitude	300 Hz Amplitude	Product	Accumulating Sum of Products
0.00083	60	60	3600	3600
0.00167	52	0	0	3600
0.00250	30	−60	−1800	1800
0.00333	52	0	0	1800
0.00417	60	60	3600	5400
0.0050	0	0	0	5400
0.0058	−60	−60	3600	9000
0.00667	−52	0	0	9000
0.0075	−30	60	−1800	7200
0.00833	−52	0	0	7200
0.00917	−60	−60	3600	10800
0.01000	0	0	0	10800

- Complex wave at fundamental frequency 100Hz, sine at 300Hz

Extracting Digital Features from Analog Sound Signals: MFCC Pipeline

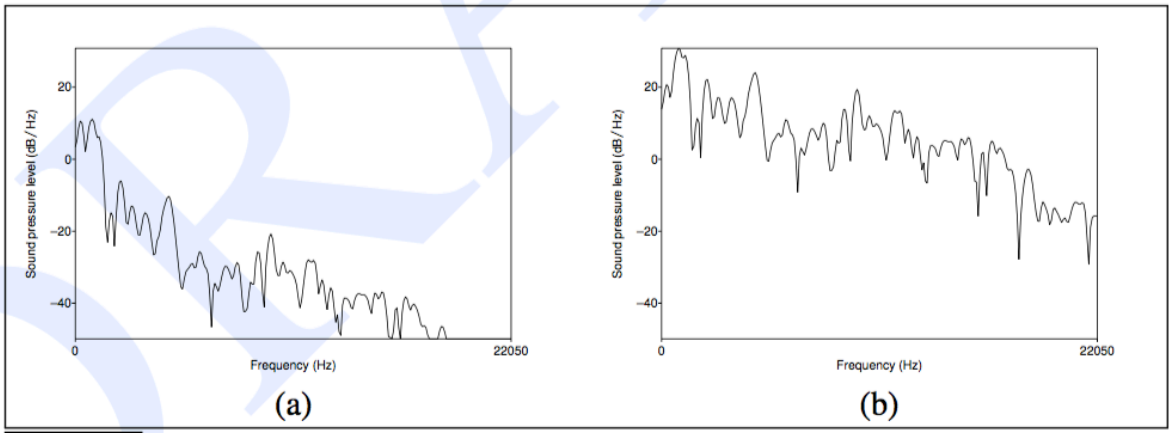


- Pipeline for extracting 39-dimensional mel frequency ceptral coefficient (MFCC) feature vector from sound signal

Analog-to-digital Conversion

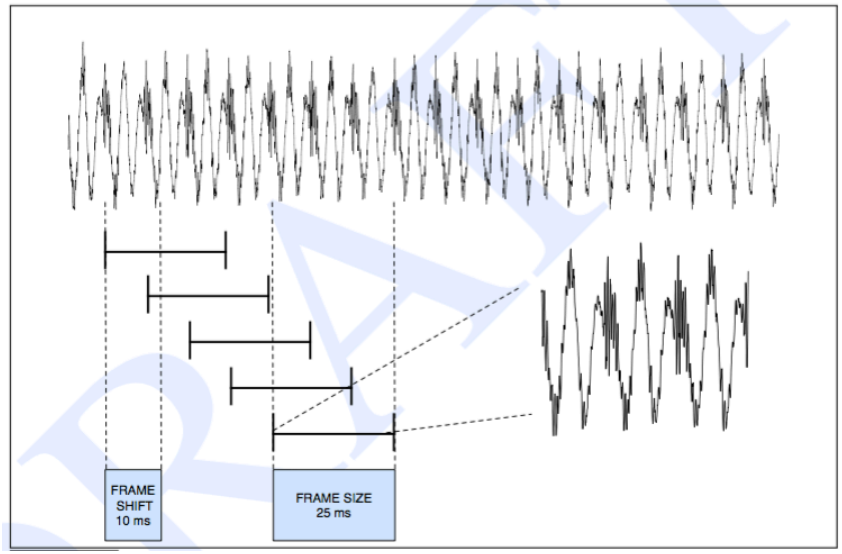
- We measure the amplitude of an analog signal at a particular time by taking a certain number of samples per second
- Knowing highest frequency component in signal, **sampling rate**, i.e., number of samples per second, has to be chosen twice as high (since at least two samples per cycle needed)
- **Nyquist frequency** is maximum frequency that can be analyzed accurately at given sampling rate (= half of sampling rate)
- Signal $x[n]$ is then digitized quantized waveform (measurements closer together than quantum size are represented identically)

Preemphasis



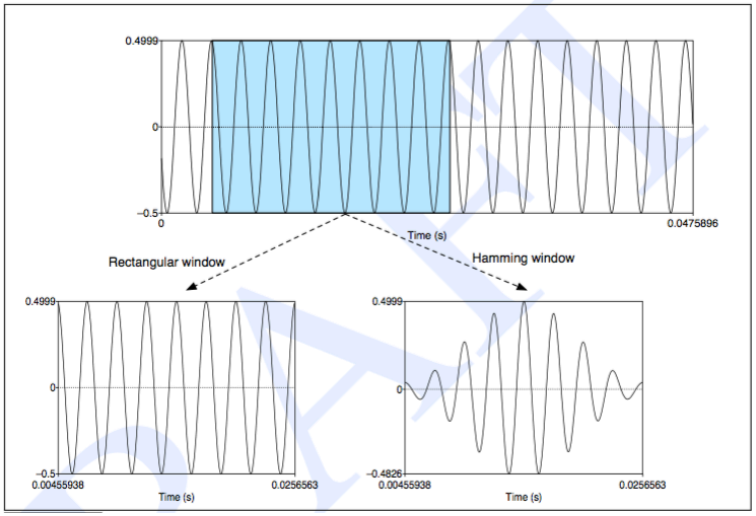
- Preemphasis **boosts energy in high frequencies** which are underrepresented in signal
- Filter equation: $y[n] = x[n] - \alpha x[n-1]$, where $0.9 \leq \alpha \leq 1.0$

Windowing



- For spectral analysis of phone, we assume that frequencies in a signal are **stationary** for a small time frame
- Signal extraction is done by multiplying signal value $y[n]$ by window value $w[n]$

Hamming Window



- Rectangular window cuts out L samples from original signal
- Hamming window shrinks values at boundaries to zero:

$$w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{L}) \text{ for } 0 \leq n \leq L-1;\ 0 \text{ otherwise}$$

Frequency Resolution, Window Length and Sampling Rate

- **Window length will determine fundamental frequency**, and thus frequency resolution in harmonics in spectral analysis

Frequency Resolution, Window Length and Sampling Rate

- **Window length will determine fundamental frequency**, and thus frequency resolution in harmonics in spectral analysis
  - Assume sampling rate of 10,000 Hz. Nyquist frequency is 5,000 Hz. Taking window of length 25.6 ms corresponds to 256 samples per cycle. Fundamental frequency of sampled wave is $10{,}000/256 = 39$ Hz.

Frequency Resolution, Window Length and Sampling Rate

- **Window length will determine fundamental frequency**, and thus frequency resolution in harmonics in spectral analysis
  - Assume sampling rate of 10,000 Hz. Nyquist frequency is 5,000 Hz. Taking window of length 25.6 ms corresponds to 256 samples per cycle. Fundamental frequency of sampled wave is $10{,}000/256 = 39$ Hz.
  - Longer windows will create smaller intervals between frequency components, but might be too long to assume stationarity
  - Lowering sample rate will have same effect, but it will also lower Nyquist frequency

DFT

- DFT
  - Look for frequency components in a complex sound wave that are multiples of the fundamental frequency
  - Determine correlation of each possible frequency component and complex wave
  - Magnitudes at these frequency bands determine spectrum of sound wave
  - Mathematics explained above
- Since DFT has complexity $\mathcal{O}(N^2)$, mostly more efficient **fast Fourier transform (FFT)** used in implementations

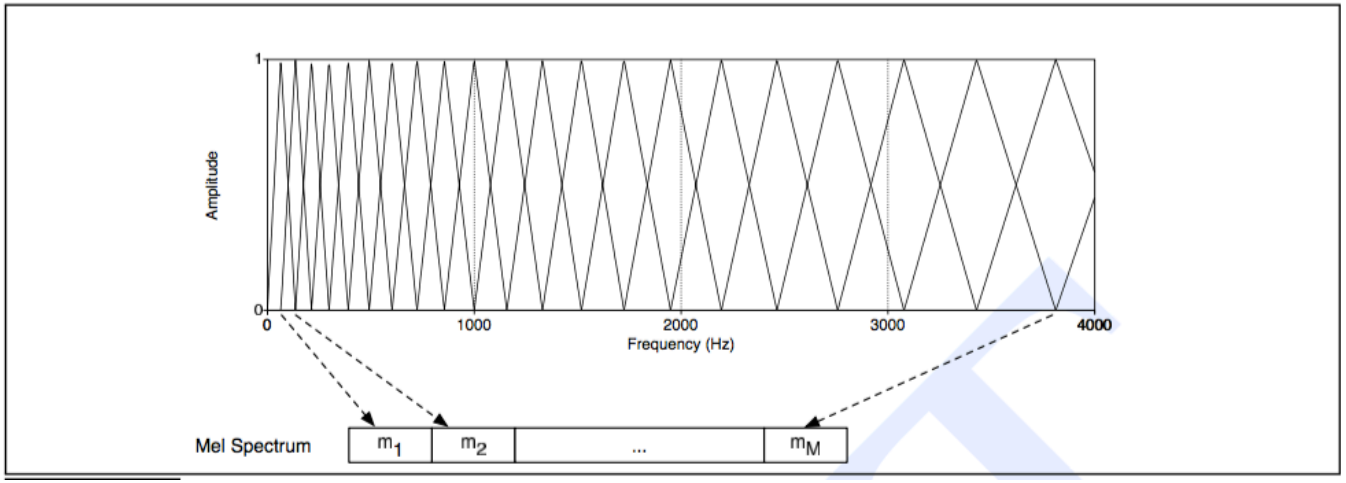
Mel Scale

- **mel** scale (as in "melody") tries to model non-linear human hearing which is less sensitive at higher frequencies, so that equidistant pitches are separated by equal mels
- mel frequency:

$$mel(f) = 1123 \ln(1 + \frac{f}{700})$$

- Linear mapping of frequency f into *mel* below 1,000 Hz, logarithmic above

Mel Filterbank



- Mel scale is realized by **triangular filters** with value 1 at center frequency, linear decrease to 0 at boundaries, spaced linearily below 1,000 Hz, logarithmically above

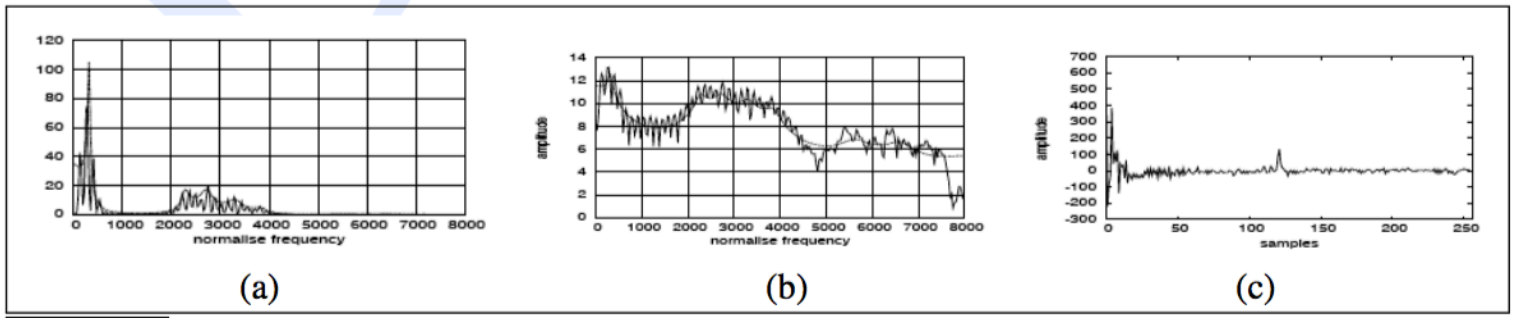
Cepstrum: Inverse DFT

- Shortcomings of spectrum generated by DFT:
  - Frequency components are highly correlated (harmonics)
  - Fundamental frequency is not important for phone detection
- Cepstrum (anagram of spectrum):
  - Spectrum of the log magnitude of the spectrum:

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi kn}{N}}\right|\right) e^{-i\frac{2\pi kn}{N}}$$

  - Intuition: Treat log-spectrum as "pseudo-signal", analyze its components

Cepstrum



- Spectrum (a), log spectrum (b), cepstrum (c)
- High frequency component in (b) caused by F0: glottal pulse
- Lower frequencies in (b) are F1, F2, etc.: vocal tract filter
- Phone detection relies of first 12 cepstral values in (c)

Deltas and Energy

- Each window is represented by **12 cepstral features**
- **Energy feature** of window: $\sum_{t=t_i}^{t_j}(x[t])^2$ from t_i to t_j
  - Helps to distinguish vowels from consonants
- For each of these 13 features, **delta** and **double delta** features for change in feature value between windows is computed
  - Helps to identify features phone properties such as stop closure and burst

Summary: MFCC features

- 12 cepstral features, 12 delta cepstral, 12 delta-delta cepstral, 1 energy, 1 energy delta, 1 energe delta-delta
- Advantages:
  - **Noise robustness:** Additive noise in non-speech regions, and average noise of microphone, can be easily detected and substracted from each frame
  - **Speaker variation:** Lower formants indicate differences like longer vocal tract in speakers, can be normalized by vocal tract length normalization
- Possible criticism:
  - DFT is linear operation, discards non-linear information
  - Decorrelation due to inverse DFT might not be necessary with deep learning (next lecture)

Visualizations

- DFT
- correlation
- complex correlation
- complete example

Acknowledgements

- Images from
  - Ladefoged (2006). Elements of Acoustic Phonetics. University of Chicago Press
  - Jurafsky & Martin (2009). Speech and Language Processing. Prentice Hall.
  - https://jackschaedler.github.io/circles-sines-signals/index.html