

© 2019 The University of Sheffield

COM3502/4502/6502 SPEECH PROCESSING

Lecture 19 Cepstral Analysis

The University Of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 1

1

© 2019 The University of Sheffield

Source-Filter Separation

- 'Cepstral analysis' is another method (*like linear prediction*) for separating the vocal tract filter response from the excitation
- It is based on the observation that the spectrum of a speech signal is the *product* of the excitation spectrum and the vocal tract frequency response

Time-domain speech signal: the vowel "iy"

Amplitude vs Time

Spectrum

dB vs Frequency

Filter →

Source →

The University Of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 2

2

© 2019 The University of Sheffield

Source-Filter Separation

- Multiplication can be converted to addition using logs ...

$$\log(a.b) = \log(a) + \log(b)$$
- Hence the '**log spectrum**' can be viewed as the summation of the log excitation spectrum and the log vocal tract frequency response

$$\log(\text{*spectrum*}) = \log(\text{*excitation spectrum*}) + \log(\text{*vocal tract frequency response*})$$

- So, if the log spectrum is treated as if it were a *waveform*, then it can be filtered!
- This approach is called '**homomorphic filtering**'

The University Of Sheffield. COM3502-4502-6502 Speech Processing: Lecture 19 slide 3

3

© 2019 The University of Sheffield

Homomorphic Filtering

The diagram illustrates the homomorphic filtering process. It starts with a plot of the 'Excitation and vocal tract spectra multiplied' (log spectrum). This signal is processed by a 'log' block. The output is then split into two paths: one through a 'Low Pass' filter followed by an 'exp' block, resulting in the 'Vocal tract spectrum'; and another through a 'Hi Pass' filter followed by an 'exp' block, resulting in the 'Excitation spectrum'.

The University Of Sheffield. COM3502-4502-6502 Speech Processing: Lecture 19 slide 4

4

© 2019 The University of Sheffield

Homomorphic Filtering

COM3502-4502-6502 Speech Processing: Lecture 19 slide 5

5

© 2019 The University of Sheffield

The Cepstrum

- As we saw in Lecture #17, a sequence can be filtered in the time domain or in the frequency domain
- This means that another way of filtering the (*log*) spectrum is to compute its *spectrum*!
- The spectrum of the (*log*) spectrum is called the '**cepstrum**' (/kepstrəm/)
- The cepstrum is defined formally as "the inverse Fourier transform of the log magnitude spectrum of a signal" ...

$$c_{\hat{n}}[m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X_{\hat{n}}(e^{j\hat{\omega}})| e^{j\hat{\omega}m} d\hat{\omega}$$

The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 6

6

© 2019 The University of Sheffield

The Cepstrum

- The cepstrum is computed in the 'quefrency' domain
- Quefrency is measured in units of *time* ($1/f$)
- Filtering in the quefrency domain is called 'liftering'

The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 7

7

© 2019 The University of Sheffield

Cepstral Analysis

The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 8

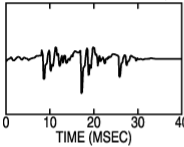
8

© 2019 The University of Sheffield

Cepstral Analysis

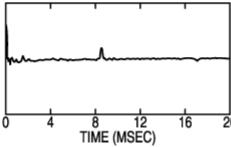
ANALYSIS FOR VOICED SPEECH

INPUT SPEECH SEGMENT
(NORMALIZED AND WEIGHTED
BY A HAMMING WINDOW)



TIME (MSEC)

CEPSTRUM

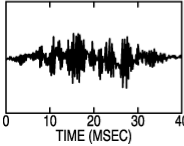


TIME (MSEC)

Schafer, & Rabiner. (1970).
System for automatic formant
analysis of voiced speech.
*Journal of the Acoustical
Society of America*, 47, 634.

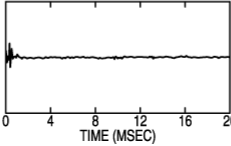
ANALYSIS FOR UNVOICED SPEECH

INPUT SPEECH SEGMENT
(NORMALIZED AND WEIGHTED
BY A HAMMING WINDOW)



TIME (MSEC)

CEPSTRUM



TIME (MSEC)

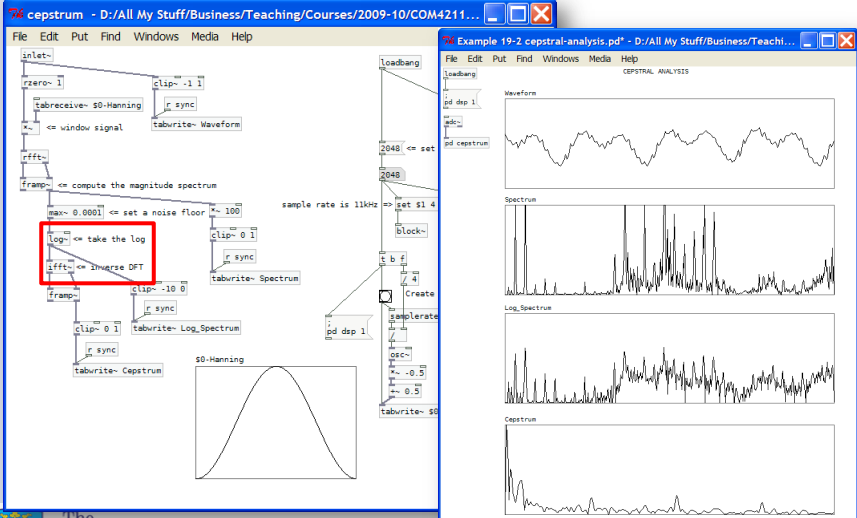
The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 9

9

© 2019 The University of Sheffield

Cepstral Analysis



The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 10

10

© 2019 The University of Sheffield

Other Cepstral Representations

- There are a number of alternatives to computing cepstral representations
- E.g. the cepstrum can be computed from the LP spectrum
- There is also an efficient method to obtain the p cepstral coefficients directly from the *linear-predictor coefficients* using the recursion ...

$$c_0 = \log G^2$$

$$c_n = -a_n - \frac{1}{n} \sum_{j=1}^{n-1} j c_j a_{n-j} \quad 0 < i \leq p$$

COM3502-4502-6502 **Speech Processing:** Lecture 19 slide 11

11

© 2019 The University of Sheffield

Applications of Cepstral Analysis

- Pitch estimation
 - based on finding the peak cepstral value in the high quefrency components
- Smoothed spectrum estimation
 - based on taking a DFT of the (*zero-padded*) low quefrency components
 - provides an alternative to the LP spectrum
- Vocoding
 - based on estimating the fundamental frequency and transmitting speech as a sequence of low quefrency cepstral frames
- Automatic speech recognition
 - based on using the low quefrency cepstral components as a representation of the (*smoothed*) spectrum
 - **'Mel Frequency Cepstrum Coefficients'** (MFCCs)

COM3502-4502-6502 **Speech Processing:** Lecture 19 slide 12

12

© 2019 The University of Sheffield

Mel-Frequency Cepstrum Coefficients

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 28(4), 357-366.

- MFCCs are a practical *approximation* to the full cepstral coefficients ...
 - the frequency scale is non-linear
 - the spectrum is computed using a filter bank
 - the cepstral coefficients are estimated using the 'discrete cosine transform' (DCT)
- Rather than implement a (*computationally expensive*) bank of digital filters, MFCCs are usually computed using a DFT whose output is grouped into 20-40 bands
- MFCCs have several attractive properties ...
 - they *decorrelate* the information in the spectrum
 - only a few parameters (10-15) are needed to represent each frame

COM3502-4502-6502 Speech Processing: Lecture 19 slide 13

13

© 2019 The University of Sheffield

The Mel Scale

- The 'Mel scale' is a *non-linear* frequency scale that reflects some aspects of auditory pitch perception
- The name derives from the word "melody"
- The scale is often regarded as being approximately linear up to 1kHz and logarithmic thereafter

$$Mel(f) = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right)$$

COM3502-4502-6502 Speech Processing: Lecture 19 slide 14

14

© 2019 The University of Sheffield

Mel Frequency Filter Bank

X

Energy in each band

The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 15

15

© 2019 The University of Sheffield

Mel Frequency Filter Bank

The University of Sheffield.

COM3502-4502-6502 Speech Processing: Lecture 19 slide 16

16

© 2019 The University of Sheffield


The Discrete Cosine Transform

- The '**DCT**' models a sequence as a sum of cosine components
- It is a *real-valued* approximation to the DFT
- It exploits the fact that the log magnitude spectrum is real-valued, symmetric with respect to θ and periodic in frequency

$$c_n = \sqrt{\frac{2}{P}} \sum_{i=1}^P m_i \cos \left[\frac{n \left(i - \frac{1}{2} \right) \pi}{P} \right]$$

... where P is the number of filterbank channels

- Note that c_0 is a measure of the signal energy



The University Of Sheffield.

COM3502-4502-6502 **Speech Processing: Lecture 19 slide 17**

17

© 2019 The University of Sheffield

Standard ASR Front-End

S
P
E
E
C
H

Fast Fourier Transform

→

Mel-Scale Filter Bank

→

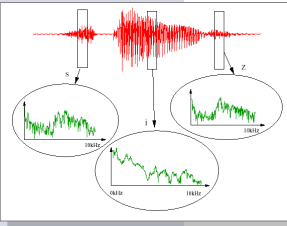
L
o
g

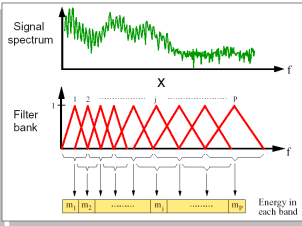
→

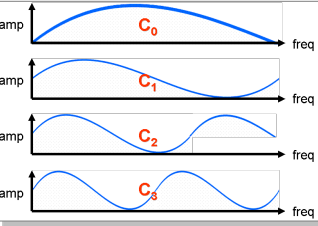
Discrete Cosine Transform


→

M
F
C
C
S










The University Of Sheffield.

COM3502-4502-6502 **Speech Processing: Lecture 19 slide 18**

18

© 2019 The University of Sheffield

Standard ASR Front-End



- 39-dimensional vector (*frame*)
 - 15-40 msec frame size
 - 10 msec frame shift
 - 100 frames per second
- Each frame contains:
 - 13 MFCCs
 - 13 1st-order derivatives (“*deltas*”)
 - 13 2nd-order derivatives (“*delta deltas*”)
- You can find out how these are used in ‘**COM4511-6511: Speech Technology**’!

The University Of Sheffield.

COM3502-4502-6502 **Speech Processing: Lecture 19 slide 19**

19

© 2019 The University of Sheffield

This lecture has covered ...

- Source-filter separation
- Homomorphic filtering
- The cepstrum
- Cepstral analysis
- The Mel frequency scale
- The Discrete Cosine Transform
- Mel frequency cepstral coefficients
- The standard ASR front-end


The University Of Sheffield.

COM3502-4502-6502 **Speech Processing: Lecture 19 slide 20**

20

© 2019 The University of Sheffield

Any Questions ?



The University of Sheffield.

COM3502-4502-6502 **Speech Processing:** *Lecture 19 slide 21*

21

© 2019 The University of Sheffield

Next time ...

Final Overview

/plʌs səm fʌn stʌf/

The University of Sheffield.

COM3502-4502-6502 **Speech Processing:** *Lecture 19 slide 22*

22