Return to Main

# LECTURE 17: SPECTRAL TRANSFORMATIONS

- Objectives:
    - Introduce perceptual linear prediction
    - Discuss speaker-dependent frequency scaling
    - Introduce vocal tract length normalization
    - Review

The original reference for perceptual linear prediction is:

> H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738--1752, 1990.

Similarly, the original reference for vocal tract length normalization is reprinted here:

> A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," *Proceedings CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

The course textbook and resource links also contain good explanations of this material.

Return to Main

**Introduction:**

**Speech Signals:**

**Signal Processing:**

**Parameterization:**

**Statistical Modeling:**

# ECE 8463: FUNDAMENTALS OF SPEECH RECOGNITION

Professor Joseph Picone
Department of Electrical and Computer Engineering
Mississippi State University

email: picone@isip.msstate.edu
phone/fax: 601-325-3149; office: 413 Simrall
URL: http://www.isip.msstate.edu/resources/courses/ece_8463

Modern speech understanding systems merge interdisciplinary technologies from Signal Processing, Pattern Recognition, Natural Language, and Linguistics into a unified statistical framework. These systems, which have applications in a wide range of signal processing problems, represent a revolution in Digital Signal Processing (DSP). Once a field dominated by vector-oriented processors and linear algebra-based mathematics, the current generation of DSP-based systems rely on sophisticated statistical models implemented using a complex software paradigm. Such systems are now capable of understanding continuous speech input for vocabularies of hundreds of thousands of words in operational environments.

In this course, we will explore the core components of modern statistically-based speech recognition systems. We will view speech recognition problem in terms of three tasks: signal modeling, network searching, and language understanding. We will conclude our discussion with an overview of state-of-the-art systems, and a review of available resources to support further research and technology development.

Tar files containing a compilation of all the notes are available. However, these files are large and will require a substantial amount of time to download. A tar file of the html version of the notes is available here. These were generated using wget:

```
wget -np -k -m http://www.isip.msstate.edu/publications/courses/ece_8463/lectures/current
```

A pdf file containing the entire set of lecture notes is available here. These were generated using Adobe Acrobat.

Questions or comments about the material presented here can be directed to help@isip.msstate.edu.

## LECTURE 17: SPECTRAL TRANSFORMATIONS

- Objectives:
    - Introduce perceptual linear prediction
    - Discuss speaker-dependent frequency scaling
    - Introduce vocal tract length normalization
    - Review

The original reference for perceptual linear prediction is:

> H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738--1752, 1990.

Similarly, the original reference for vocal tract length normalization is reprinted here:

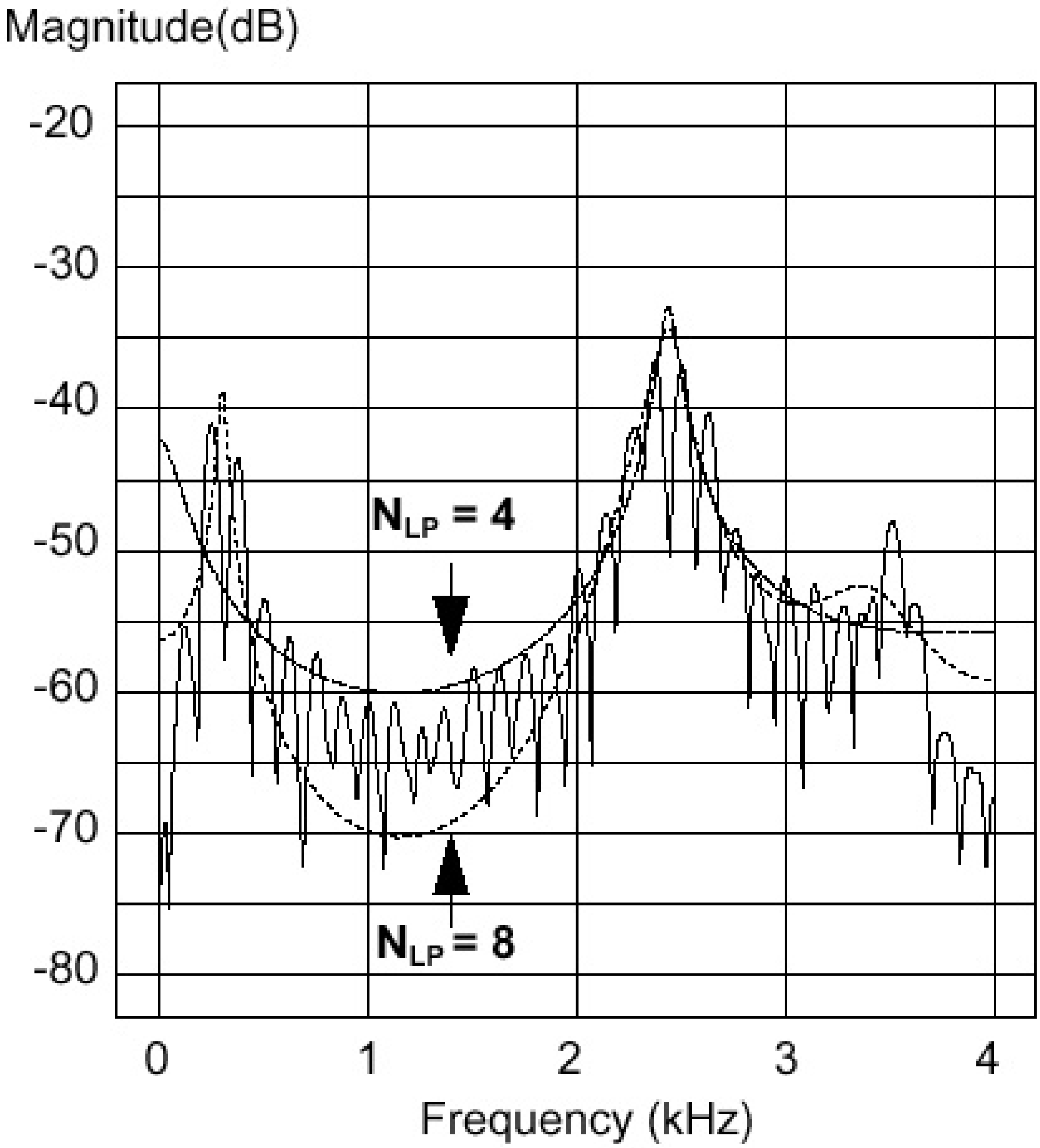
> A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," *Proceedings CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

The course textbook and resource links also contain good explanations of this material.

## SPECTRAL MATCHING INTERPRETATION

Recall that the LP model uses mean-square error approach to optimize its coefficients. This implies:

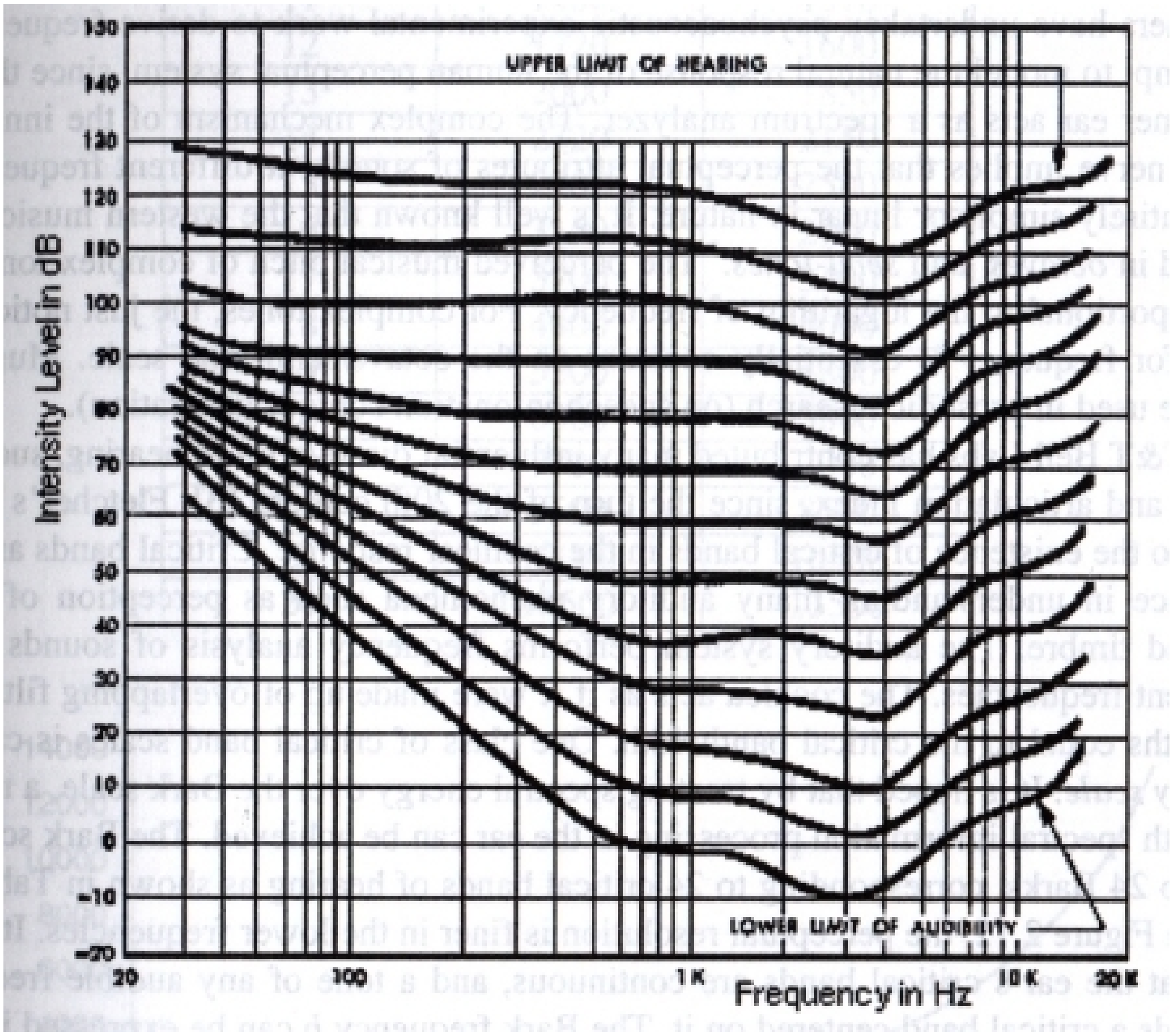
- The LP model attempts spectral flatten the error signal.
- The LP model focuses on the extremely high or low energy areas of the spectrum - whatever it takes to makes the error signal spectrum as flat as possible. Example:



- Note that the eighth-order analysis models the floor of the spectrum more precisely than the third formant.

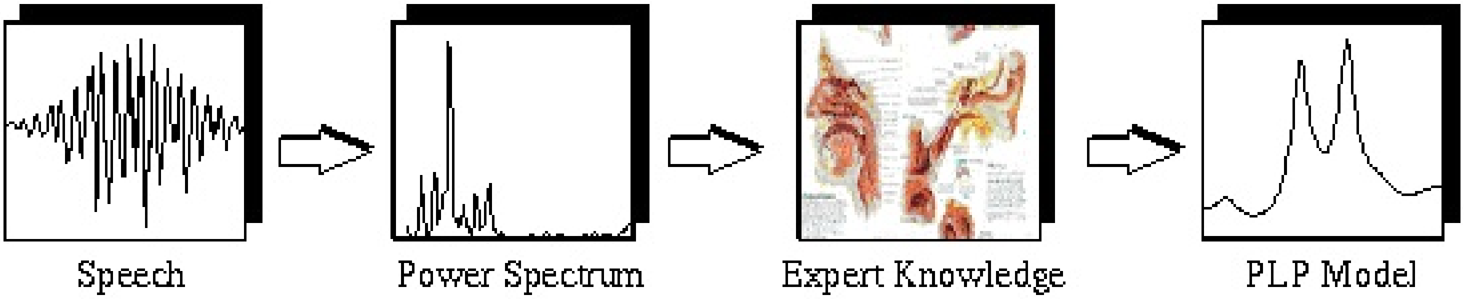
## EQUAL LOUDNESS CURVES

Recall our observation that perceptual loudness of a sound is a function of its absolute intensity:



- The sensitivity of the ear varies with the frequency content and the quality of a sound.
- The graph above represents equal loudness contours adopted by the ISO (ISO 226).
- Hearing sensitivity peaks at 4K Hz, and has a secondary peak at 13K Hz.

# PERCEPTUAL LINEAR PREDICTION

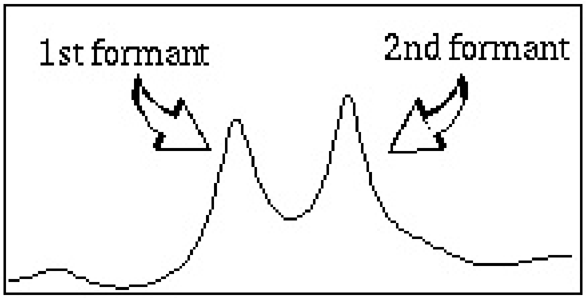


- Psychophysical concepts:

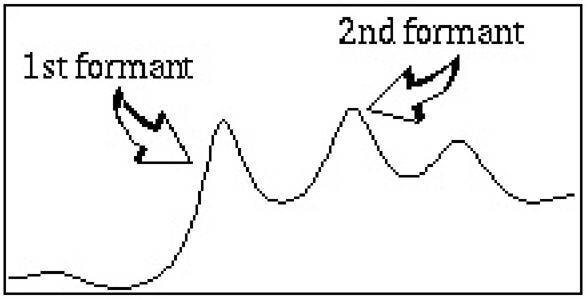
    - Critical-band spectral resolution,
    - Equal loudness curve,
    - Intensity-loudness power law.

- PLP coefficients still model the important frequencies in noise.

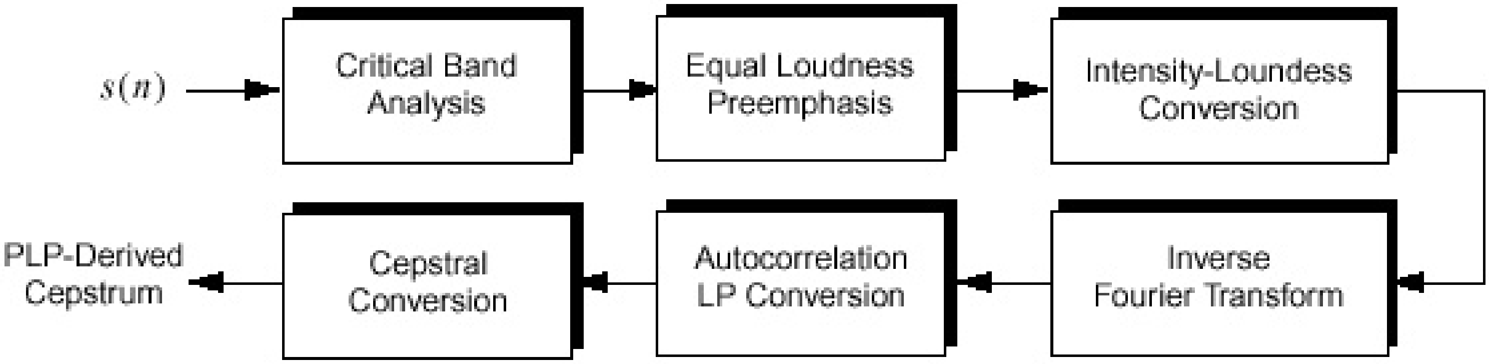
Clean Speech (25.3 dB SNR)



Noisy Speech (2.1 dB SNR)



# PERCEPTUAL LINEAR PREDICTION
# BLOCK DIAGRAM


$s(n)$
Critical Band Analysis
Equal Loudness Preemphasis
Intensity-Loundess Conversion
Inverse Fourier Transform
Autocorrelation LP Conversion
Cepstral Conversion
PLP-Derived Cepstrum

- Goals:
  - Apply greater weight to perceptually-important portions of the spectrum
  - Avoid uniform weighting across the frequency band
- Algorithm:
  - Compute the spectrum via a DFT
  - Warp the spectrum along the Bark frequency scale
  - Convolve the warped spectrum with the power spectrum of the simulated critical band masking curve and downsample (to typically 18 spectral samples)
  - Preemphasize by the simulated equal-loudness curve:
    - Simulate the nonlinear relationship between intensity and perceived loudness by performing a cubic-root amplitude compression
  - Compute an LP model
  - Compute an LP-derived cepstrum
- Claims:
  - Improved speaker independent recognition performance
  - Increased robustness to noise, variations in the channel, and microphones

# EQUAL LOUNDNESS PREEMPHASIS AND PERCEIVED INTENSITY

- Equal loudness preemphasis is implemented using:

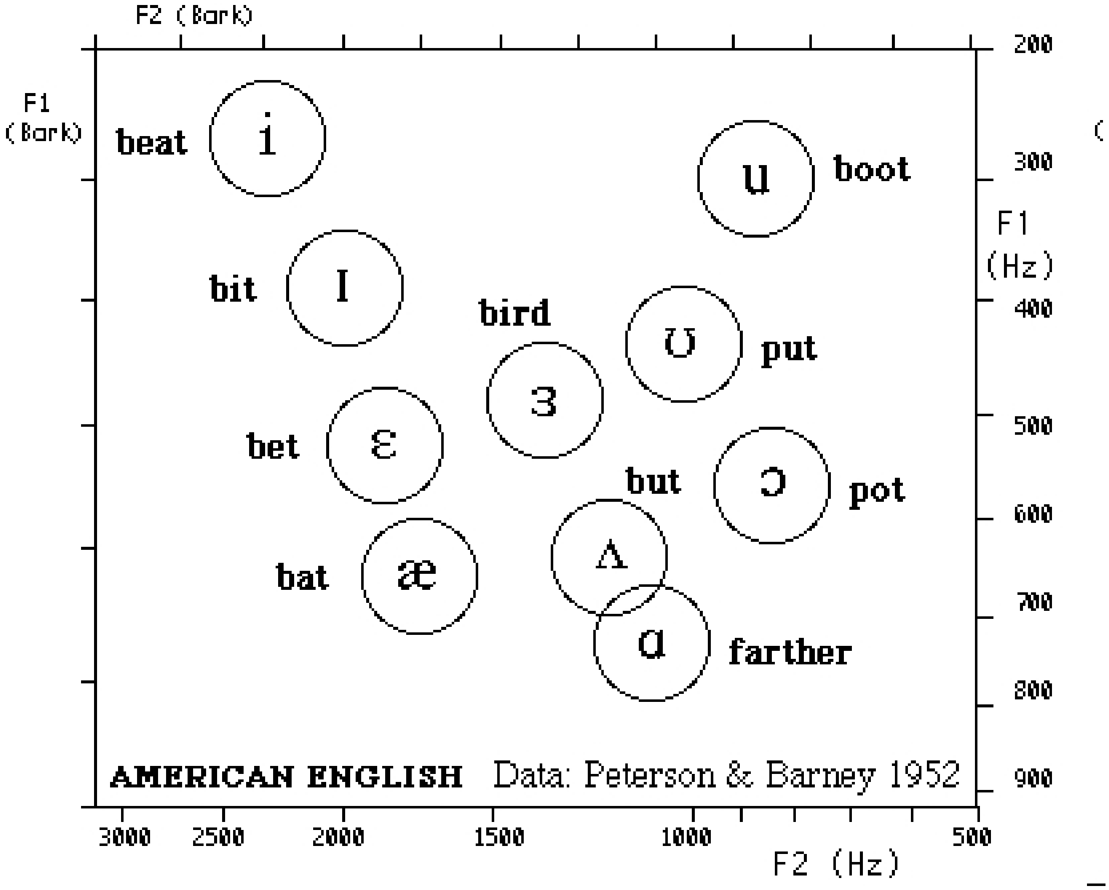
$$E(\omega) = \frac{(\omega^2 + 56.8\text{x}10^6)\omega^4}{(\omega^2 + 6.3\text{x}10^6)(\omega^2 + 0.38\text{x}10^9)(\omega^6 + 9.58\text{x}10^{26})}$$

- Perceived loudness can be approximated using:

$$L(\omega) = I(\omega)^{1/3}$$

## FORMANT FREQUENCY DISTRIBUTIONS FOR VOWELS

- Recall the Peterson-Barney vowel data:



Mean formant values of 33 male speakers for ten American English vowels. (Data from: G. Peterson and H.L.Barney, "Control Methods Used In A Study Of The Vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175-184, 1952. Figure from: Projektit: Vowel Charts.)

- What causes this natural variation?

- Is this type of variation desirable?

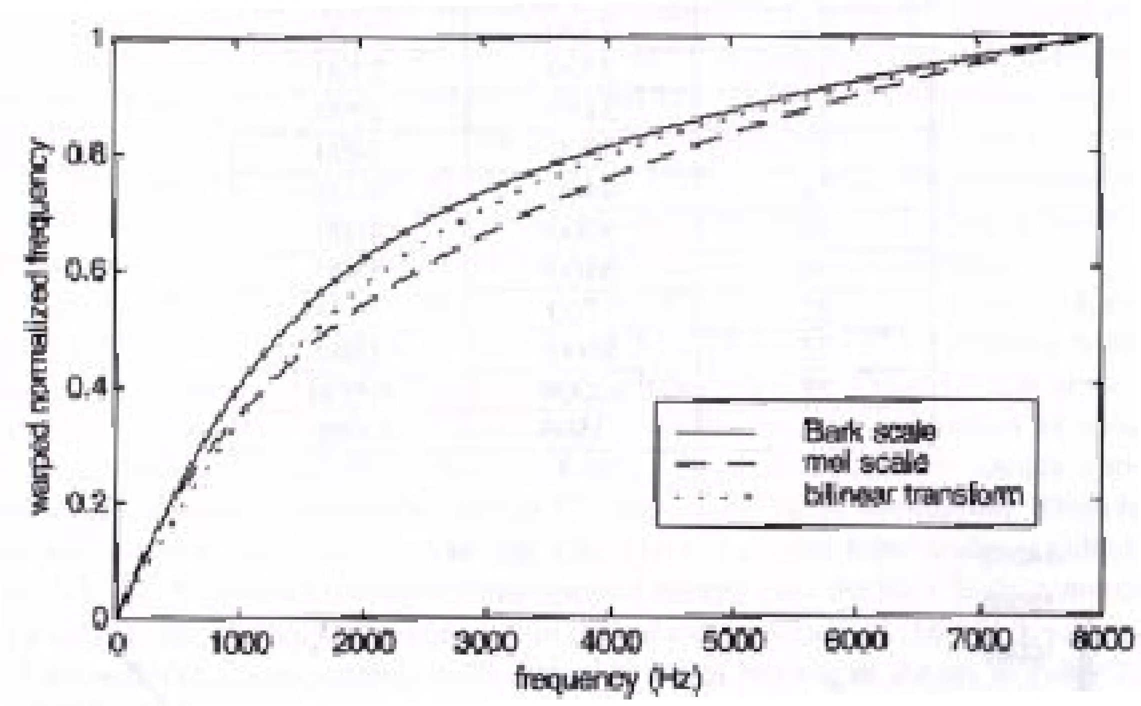
- How can we offset this variation?

## SPEAKER-DEPENDENT FREQUENCY WARPING

- Recall the bilinear transform:

$$s = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \qquad \Omega = \omega + 2\arctan\left(\frac{\alpha\sin(\omega)}{1 - \alpha\cos(\omega)}\right)$$

- This was a good approximation to the Bark and mel scales:



- How can we compute the optimal warping factor?

## DIRECT MEL-SCALE FREQUENCY WARPING

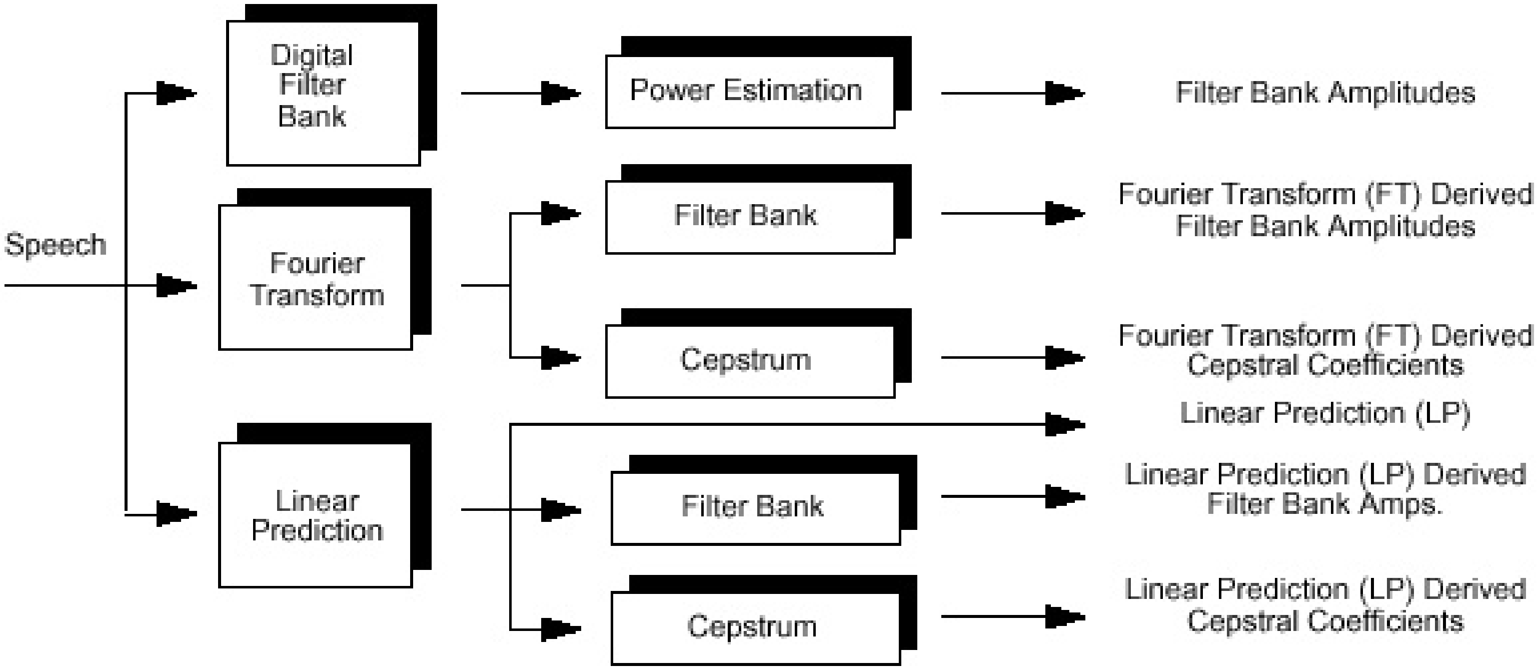
- We can warp the linear frequency axis directly. Let $k\Delta f_{mel},\ k = 1, \ldots, K$ denote the center frequencies on the mel scale. We can warp these frequencies using a simple linear transformation:

$$f^{\alpha}_{Hz}(k\Delta f_{mel}) = 700(10^{(k\Delta f_{mel})/2595} - 1)/\alpha$$

- We can also warp the discrete Fourier transform samples directly using a similar linear compression.
- A typical range for the warping factors is [0.8, 1.2].
- What are the relative merits of this approach?

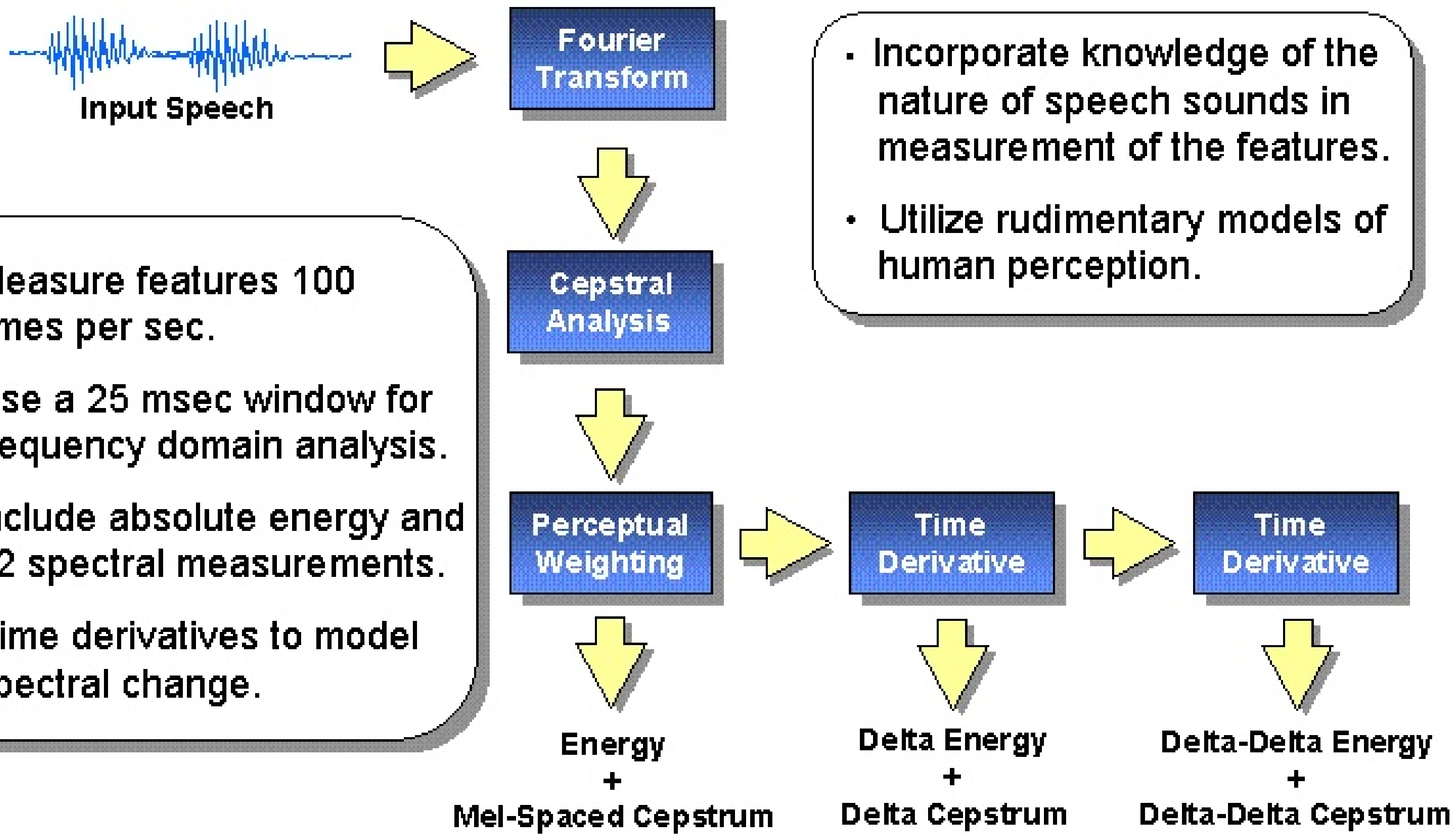
# ALTERNATIVE METHODS FOR FREQUENCY DOMAIN ANALYSIS

We have now established two different ways to perform a filterbank analysis of the speech signal (temporal and spectral):


Speech
Digital Filter Bank
Fourier Transform
Linear Prediction
Power Estimation
Filter Bank
Cepstrum
Filter Bank
Cepstrum
Filter Bank Amplitudes
Fourier Transform (FT) Derived Filter Bank Amplitudes
Fourier Transform (FT) Derived Cepstral Coefficients
Linear Prediction (LP)
Linear Prediction (LP) Derived Filter Bank Amps.
Linear Prediction (LP) Derived Cepstral Coefficients

The most popular front ends are those that use cepstral coefficients dervied from the Fourier transform. Why?

## A TYPICAL SPEECH RECOGNITION FRONT END



Input Speech
Fourier Transform
Cepstral Analysis
Perceptual Weighting
Time Derivative
Time Derivative
Energy + Mel-Spaced Cepstrum
Delta Energy + Delta Cepstrum
Delta-Delta Energy + Delta-Delta Cepstrum

- Incorporate knowledge of the nature of speech sounds in measurement of the features.
- Utilize rudimentary models of human perception.

- Measure features 100 times per sec.
- Use a 25 msec window for frequency domain analysis.
- Include absolute energy and 12 spectral measurements.
- Time derivatives to model spectral change.

# Perceptual Linear Prediction

A combination of DFT and LP techniques is perceptual linear prediction (PLP) [10].

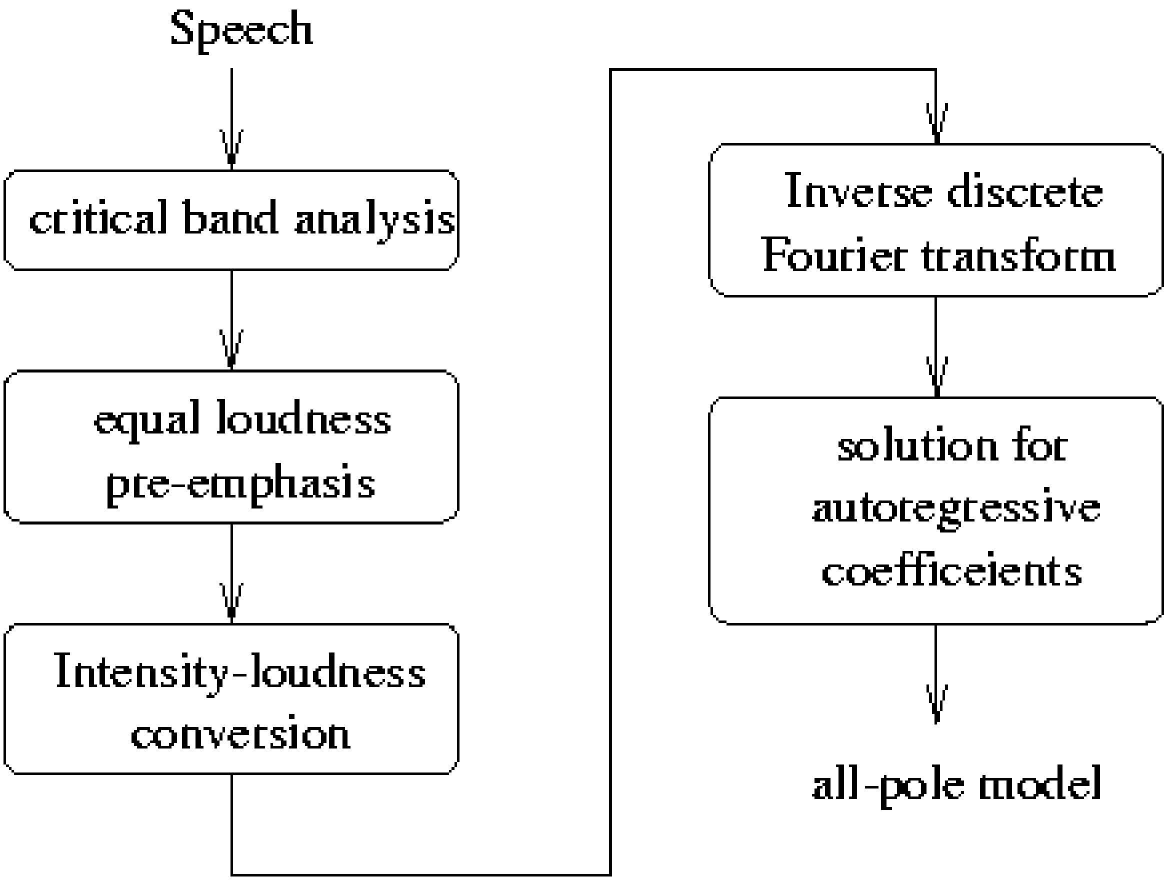

Speech
critical band analysis
equal loudness pre-emphasis
Intensity-loudness conversion
Inverse discrete Fourier transform
solution for autoregressive coefficeients
all-pole model

**Figure 41:** Perceptual Linear Prediction

### 2.3.2 Perceptual Linear Prediction Analysis

The perceptual linear prediction analysis (PLP) is a combination of spectral analysis and linear prediction analysis [8]. PLP technique uses concepts from the psychophysics of hearing to compute a simple auditory spectrum. Figure 2.4 illustrates the components of the PLP analyses.

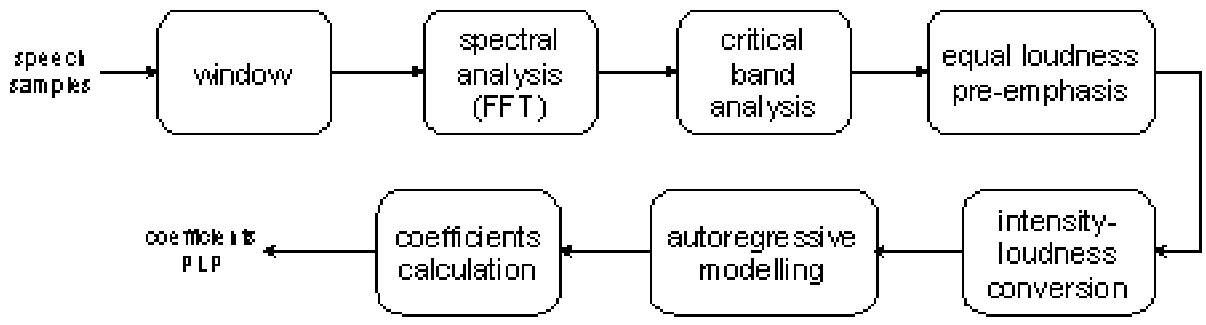


Fig. 2.4: Perceptual linear prediction analysis (PLP)

After the speech samples are weighted by a window (e.g., Hamming-window) and transformed into the frequency domain (usually by short term FFT) they are converted to a power spectrum. This spectrum is warped into a Bark scale using the approximation:

$$\Omega(\omega) = 6\cdot\ln\left(\frac{\omega}{1200\cdot\pi}+\sqrt{\left(\frac{\omega}{1200\cdot\pi}\right)^2+1}\right) \qquad (2.3)$$

where &omega is the angular frequency in rad/s and $\Omega$ represents the Bark frequency. The advantage of that transformation is to mimic the human earring process in frequency groups.

The Bark scaled spectra is merged with the power spectra of the critical band filters. This simulates the frequency resolution of the ear, which is approximately constant on the Bark scale. The resulted samples of the critical band power spectrum with the approximation of the critical band curve $\Psi(\Omega)$ can be written as follows:

$$\Theta(\Omega_i) = \sum_{\Omega}\left(P(\Omega-\Omega_i)\cdot\Psi(\Omega)\right) \qquad (2.4)$$

The equal loudness pre-emphasis is done in order to compensate the non-equal perception of loudness at different frequencies and simulates the sensitivity of hearing about the 40 dB level. If $E(\Omega)$ is an approximation to the non-equal sensitivity of human hearing, the equal loudness pre-emphasis can be written as:

$$E(\Omega(\omega)) = E(\omega)\cdot\Theta(\Omega(\omega)) \qquad (2.5)$$

Except very loud or very quiet sounds, the perceived loudness $\Gamma(\Omega)$ is approximately the cube root of the intensity. This is well known as the power law of hearing and simulates the non-linear relation between the intensity of sounds and its perceived loudness.

$$\Gamma(\Omega) = \sqrt[3]{E(\Omega)} \qquad (2.6)$$

The equal loudness pre-emphasis and the intensity loudness conversion reduce the spectral amplitude variation of the critical band spectrum. To obtain coefficients, an all-pole model has to be solved with the help of the autocorrelation method [8, 10]. The resulted autoregressive coefficients can be transformed into some other sets of parameters of interest.

top

2.3 Feature Extraction (2.3.1 Cepstral Analysis) 2.3.2 2.3.3 Relative Spectral Analysis

context