EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 29: April 3, 1995

Lecturer: Prof. Nelson Morgan *Scribe: William Chen*

29.1 Where we are now

So far, we have covered the basics of audio signal processing. The first portion of the course was devoted to understanding how humans produce and perceive speech. Understanding how nature does this is a science that can lead to useful models and stimulate new technology. The second portion of the course is to apply this knowledge to making machines synthesize and recognize speech.

Machine synthesis of speech, or synthesizers, were covered in lectures 23-27 and described the vocoder and glovetalkII systems. This lecture begins machine recognition of speech, or ASR, which will cover the following areas:

- Feature Extraction
- (Statistical) Pattern Recognition
- Sequence Recognition
- Complete Systems

Ultimately, we want to understand how these different ideas are used in current automatic speech recognition systems.

29.2 Introduction to Pattern Recognition

The problem of Automatic Speech Recognition actually belongs to a more general class of problems, known as pattern recognition. The idea is that given some data, how can we extract the relevant information and then classify it [DH73].

The classical pattern recognition view of speech is shown in Figure 29.1.

Speech is corrupted by additive and convolutional noise. The role of the feature extractor is to clean the speech up and then represent it with a small number of variables that are useful for discriminating between them. The variables form a feature vector, and the vector can be thought of as embedded in a euclidean space. The classifier then divides this space up according to some criterion and determines where each incoming feature vector belongs.

29.3 Feature Extraction

As Figure 29.2 shows, feature extraction for ASR starts by windowing the incoming speech signal. Several different types of windowing functions are commonly used, such as the Rectangular window and the Hamming

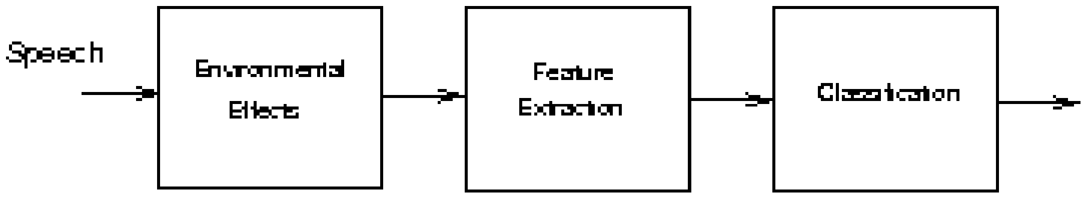

Figure 29.1: Classic View

window. The basic tradeoffs are more clearly illustrated in the Fourier Transform domain. The first tradeoff is window length. We would like a short window so that the speech is relatively stationary. However, a short window has a wide main lobe, which smears the frequencies of the speech signal and thus gives poor frequency resolution. The second tradeoff is the shape of the window. A rectangular window is simple because it just multiplies the speech signal by 1 or 0. But this simple window has high sidelobes in the frequency domain and thus distorts the neighboring frequencies, known as "leakage".

Once the windowing is done, several signal processing techniques can be used to extract the feature vector:

- Filter Banks
- LPC
- Mel Cepstrum and PLP
- Delta Cepstrum
- Robustness - Cepstral Mean Subtraction and RASTA
- Radical Front Ends - Auditory Models

29.4 Standard Feature Extraction

The first two methods of feature extraction incorporate typical signal and stochastic processing techniques to a speech segment [SR74] [M75].

29.4.1 Filter Banks

A filter bank front end divides up a speech segment according to frequency content. The speech signal is sent through a bank of bandpass filters, each of which passes the speech within a certain frequency band, and features from each filter are then combined to form the feature vector. Typically the bandpass filters have been spaced according to some perceptual criterion, such as mel or bark spacing.

Advantages: No strong assumptions have been made about the speech signal. Disadvantages: a) Since each filter produces a set of variables, there will be a relatively large number of variables to compute. b) Also there is no built in discrimination for the large variability of real speech.

29.4.2 Linear Prediction Coding

An LPC front end basically tries to model a segment of speech in the time domain [M75]. LPC predicts the next sample of speech based on a linear combination of the previous samples. Since a linear combination is used, LPC can be implemented by an all-pole filter. The values for the filter coefficients are determined by solving a system of equations that minimizes the mean squared error between the speech signal and the predicted speech signal. The error signal and the filter coefficients are used as the feature vector.

Advantages: a) Typically less computation than a filter bank or FFT. b) Converts to cepstral coefficients which are roughly orthogonal c) Includes speaker characteristics good for speaker ID.

29.5 Perceptually based Feature Extraction

Both Mel Cepstrum and PLP are very similar [H90]. They model the power spectrum of a segment of speech and then perform further processing on the spectrum to reduce the amount of information but in a way that minimizes perceptual distortion.

29.5.1 Mel Cepstrum and PLP

Both methods take the spectrum of speech and shape the information content based on a perceptual metric. Once this is done, cepstral coefficients are derived for the feature vector. The derivation of cepstral coefficients belongs to the class of nonlinear filtering. In general, cepstral analysis turns time-domain convolution into addition which allows easy separation of relevant speech parameters, ie pitch and voicing. The details of both methods are shown in Figure 29.3 and PLP is explained below:

- A power spectrum for the speech segment is first computed using straight FFT.
- The power spectrum is integrated within critical bands and resampled. Two things here to pay attention to: the integration reduces the amount of info and the choice of critical bands are based on the variation of the ear's critical bandwidths.
- Next, the spectrum is weighted to emphasize those frequencies where the ear is the most sensitive and to demphasize those frequencies where the ear is less sensitive.
- A cubic root is applied to the spectrum to mimic the power-law of hearing.
- Finally the inverse DFT is taken to derive cepstral coefficients.

Mel Cepstrum accomplishes the same overall processing although the exact steps are quite different.

Advantages: a) Less affected by speaker variability. b) Most recognizers integrate some form of this in the front end.

29.6 Robust Feature Extraction

Previous feature extraction techniques relied heavily on the power spectrum and cepstrum coefficients of speech. However, the absolute amplitude of the power spectrum/cepstrum coefficents is not always an accurate feature for pattern recognition since the amplitude and shape changes with a simple turn of the head from a microphone. The next few methods show simple ways of providing extra robustness to the features.

29.6.1 Delta Cepstrum

Delta Cepstrum looks at the local dynamics of speech [F91]. The notion here is that sound perception depends on spectral difference. So the Delta Cepstrum calculates the spectral difference by looking at the current and previous segments of speech to compute the time derivatives of the cepstrum. The relationship between spectral difference and cepstral difference is given below:

$$\frac{d}{dt} log(S(\omega, t)) = \sum_{n=-L}^{L} \frac{d}{dt} C_n(t) e^{jn\omega}$$

Some systems use just the delta cepstrum in the feature vector while others use both the cepstrum and delta cepstrum in the feature vector.

Advantages: a) Robustness to Spectral Slope b) Incorporates a temporal window. Disadvantage: a) Dismisses too much information as it attenuates perceptually important information in the 1 to 10 Hz frequency range. b) Resulting features don't look like ordinary spectra because the lower modulation spectra is gone.

29.6.2 CMS and Rasta

The cepstral mean subtraction is similar to Delta Cepstrum in the sense that a difference is also taken to get rid of slow varying convolutional noise. But here the view is global, since the cepstral mean is subtracted from the entire utterance, not just a few of the previous frames of speech. This gets rid of slow varying convolutional noise as follows:

$$\begin{aligned} x(t) &= s(t) * h(t) \\ |X(f)|^2 &= |S(f)|^2 |H(f)|^2 \\ log|X(f)|^2 &= log|S(f)|^2 + log|H(f)|^2 \end{aligned}$$

If the convolutional noise, h(t), is slow varying, then log—H(f)— is pretty much a constant and can be easily subtracted.

RASTA is the online solution to CMS, since CMS is not feasible for real-time implementations [HM94]. The procedures for RASTA are shown below in Figure 29.4.

29.6.3 Comparison of RASTA, CMS, and Delta Cepstrum

Actually all three procedures are pretty similar. In fact RASTA and Delta Cepstrum are both implemented in a similar fashion as described in Figure 29.4. Note the bandpass filters. They both severely attenuate dc and low frequency components. The Delta Cepstrum, however, has a very narrow bandpass filter as shown in Figure 29.5 and does not implement the Bank of Expanding Static Nonlinearities in Figure 29.4.

Once again RASTA is just the online implementation of CMS. Whereas as CMS just took out the DC term of a short term log spectrum, RASTA takes out this term but, like Delta Cepstrum, emphasizes spectral change.

29.6.4 Auditory Models

Below is a list of the auditory models mentioned in class.

- Seneff Model
- Ghitza
- Holton
- others

Most of them have interesting properties that save computation. Basically they can be divided up into two groups- envelope and synchrony extraction as shown in Figure 29.6.

References

[DH73] Duda and Hart, **Pattern Recognition and Scene Analysis**, Wiley-Interscience, 1973.

[F91] S. Furui, "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics", Proceedings of ICASSP '86, pp. 1991-1994, April, 1986.

[H54] J. Makhoul, "Linear Prediction: A Tutorial Review", **Proceedings of the IEEE**, pp. 561-580, April, 1975.

[H90] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech", The Journal of the Acoustical Society of America, pp. 1738-1752, April, 1990.

[HM94] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, pp. 578-589, vol. 2, no. 4, October, 1994.

[SR74] R. Schafer and L. Rabiner, "Digital Representations of Speech Signals", **Readings in Speech Recognition**, Waibel and Lee, pp. 49-64, 1960.

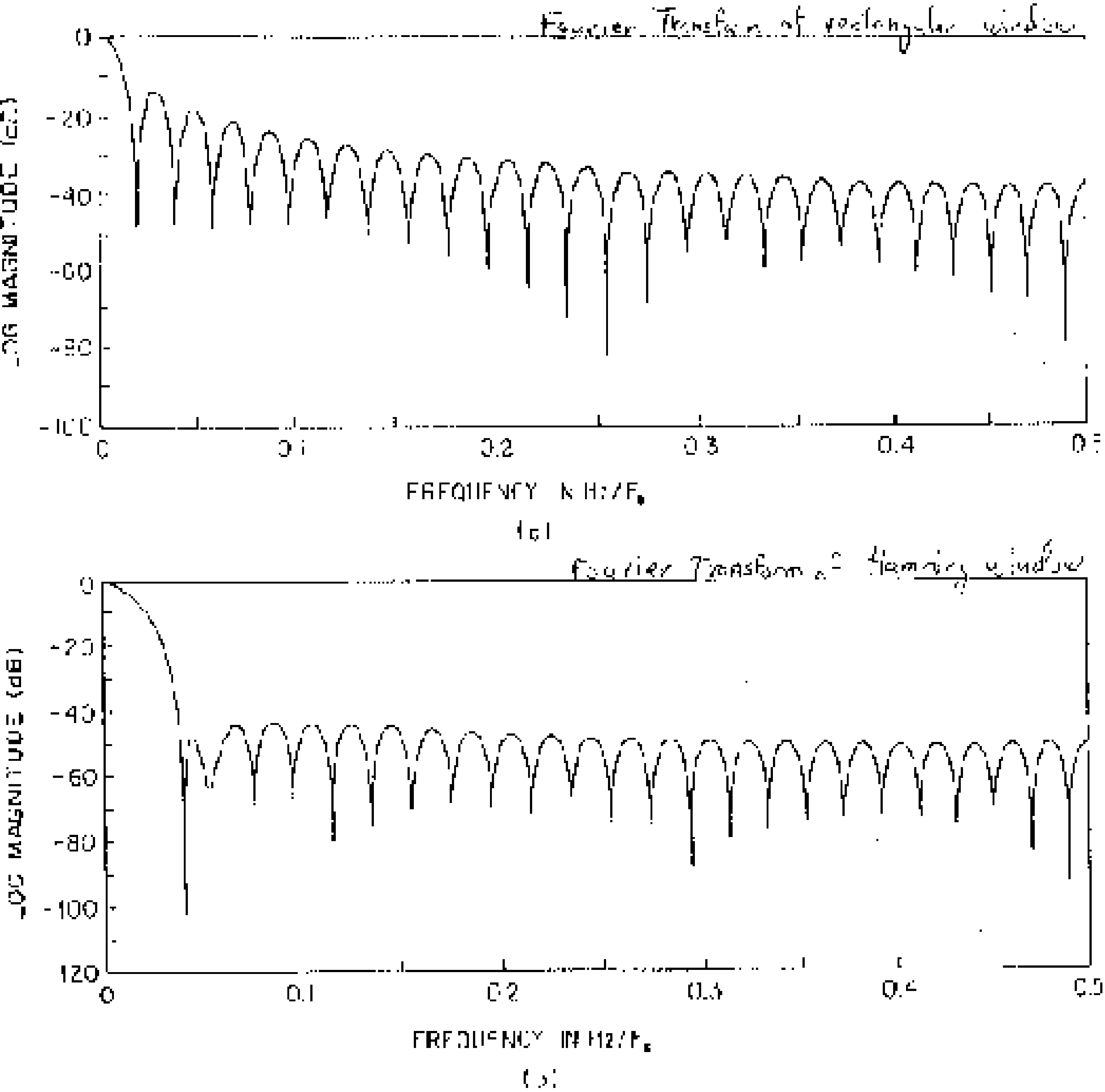

Figure 29.2: Comparison of the Fourier Transform of Two Popular Windows

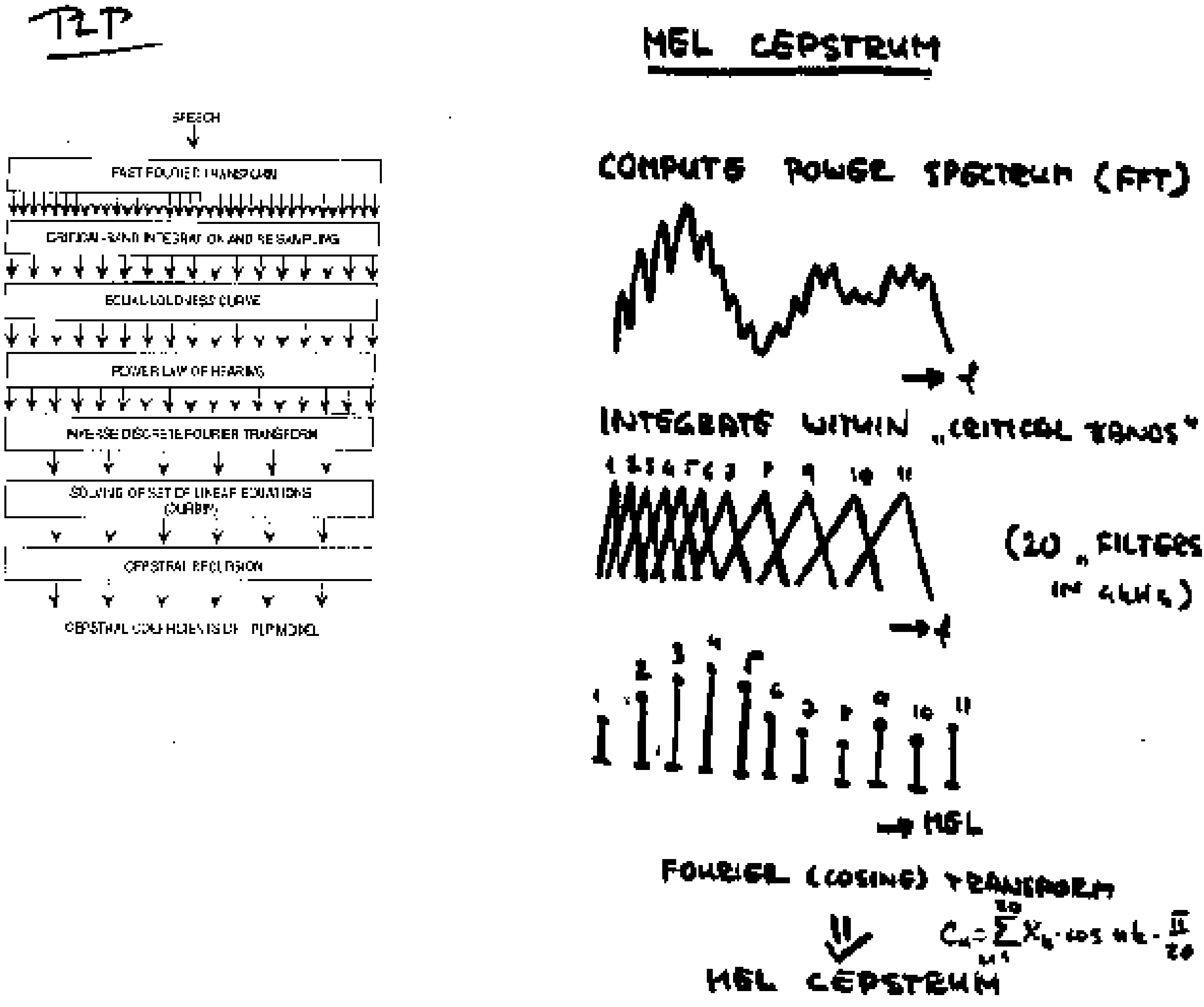

Figure 29.3: Mel-Cepstrum vs PLP

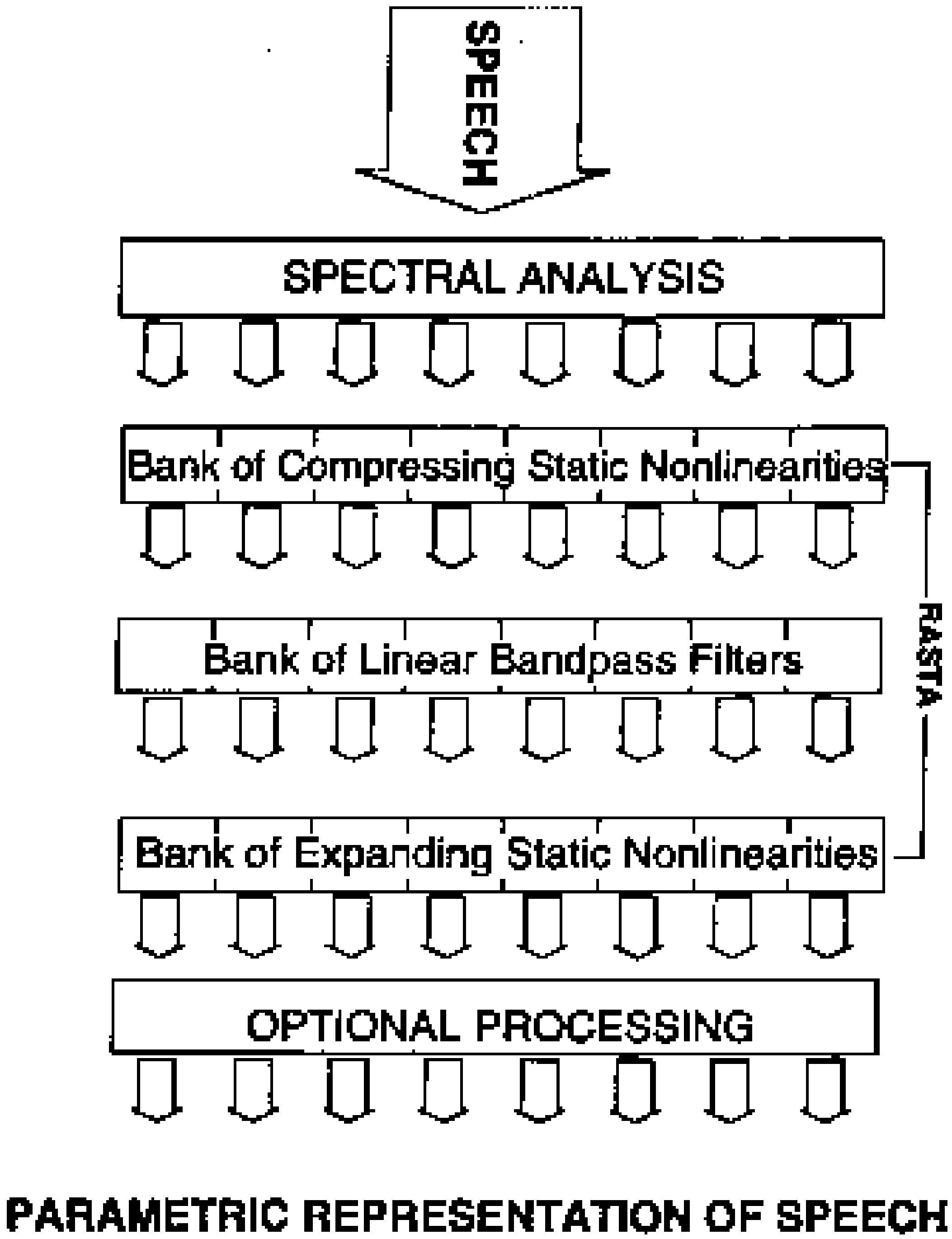

Figure 29.4: RASTA

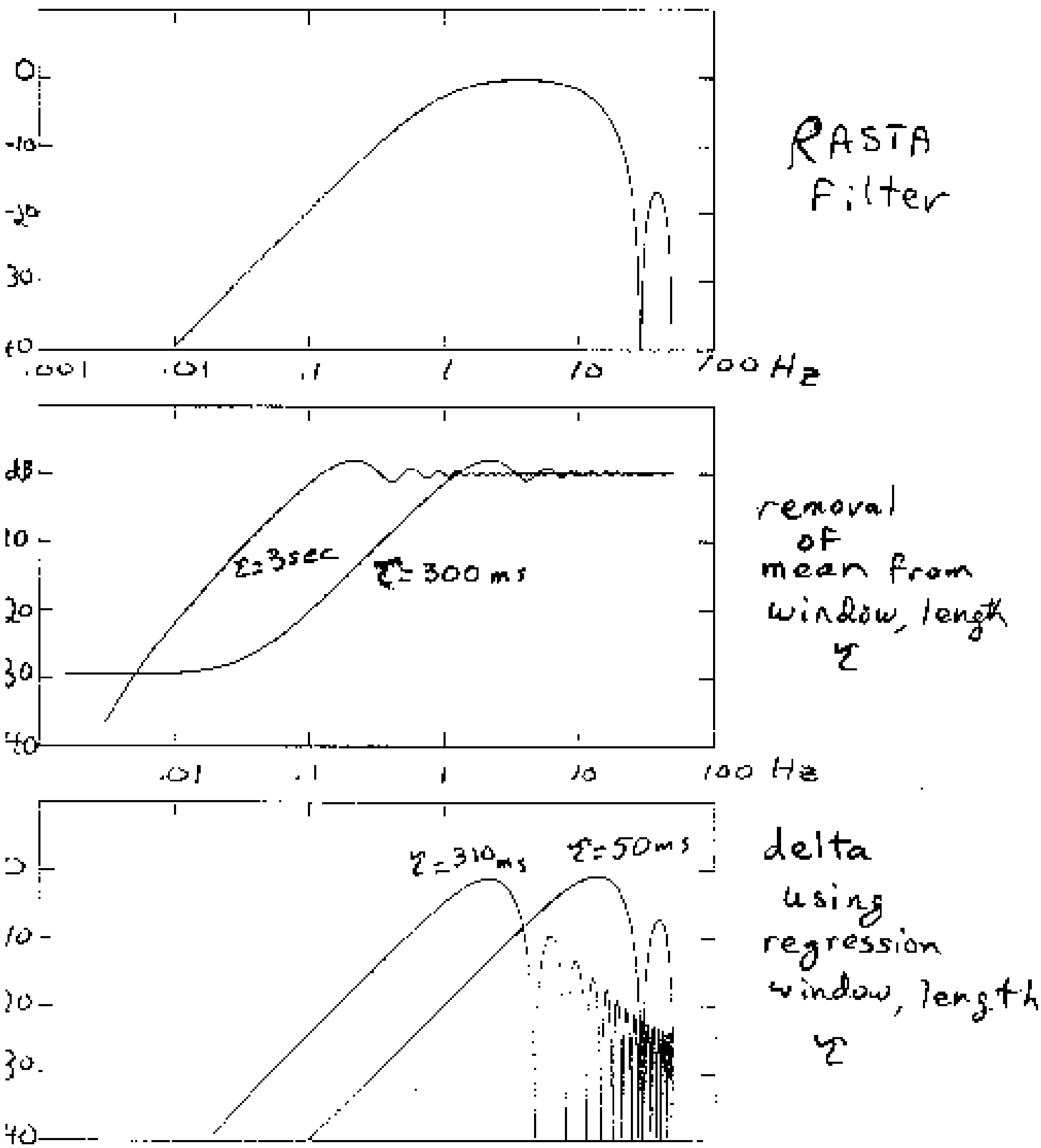

Figure 29.5: Comparison of RASTA and Delta Cepstrum Bandpass Filters

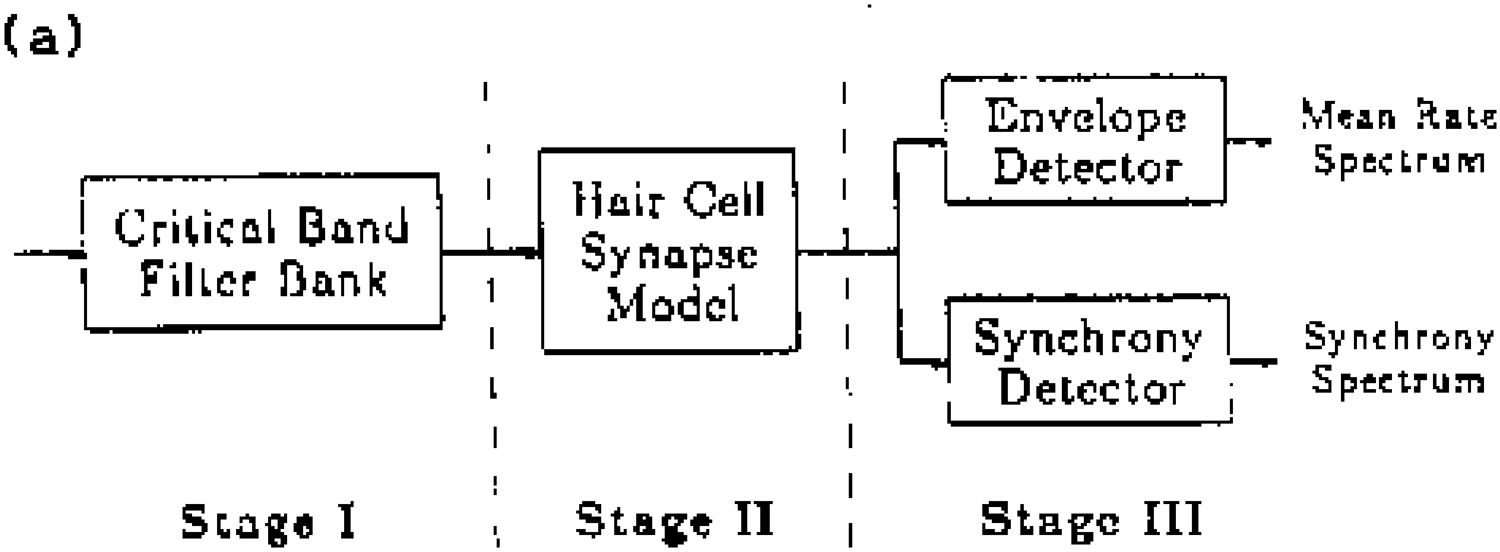

Figure 29.6: Auditory Models