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

Lecture 23: March 13, 1995

Lecturer: Prof. Ben Gold *Scribe: Jeff Gilbert*

Channel Vocoders

23.1 Analysis-Synthesis Concept

The process of coding or compressing speech can be broken into an *analysis* part and a *synthesis* part. The analysis part is akin to humans' perception of speech. The speech is broken down into some more basic, less redundant form. An example of very complete compression would be to break speech down into words or text. Text is a much more compact (yet less informative) representation than the actual speech waveform.

The synthesis side then has to reconstruct the original signal from the condensed form. For the previous example, a human could do this by reading the text.

In compression, there is always a tradeoff between reducing the bandwidth required to transmit the signal and losing relevant details which will then impair intelligibility or omit information. In the above example, much compression was achieved but at the expense of losing the identity of the speaker as well as all other supra-text information such as mood and emphasis.

23.1.1 Human Analysis-Synthesis

One example of a human speech analysis / synthesis system would be a *mimic*. A mimic hears someone else speak and tries to replicate the sounds. The analysis begins in the mimic's cochlea. The cochlea acts like a filter bank with wider filters at higher frequencies. As we will see, a vocoder's analysis often begins with a similar set of filter banks. In the case of the human mimic, the auditory cortex and other parts of the brain performs further analysis. Then the sounds are resynthesized by the vocal chords and vocal tract.

The analysis performed by a vocoder usually stops at the equivalent of the auditory nerve – beyond that we do not know how to perform further analysis.

23.2 Separation of Pitch and Spectral Envelope

A key concept in both human and machine perception (especially as it applies to speech coding) is the separation of the pitch information from the spectral envelope. (See Figure 23.1) This allows for better compression both because each can be transmitted using different precision and also each of them change more slowly alone than their product changes.

If you consider speech in terms of the source / filter model, with the source (either vocal chords or turbulence in the vocal tract) contributing to the fundamental frequency and the vocal tract controlling the spectral shape, separating these two elements of the signal seems like a good idea since they are more or less independent.

23.2.1 What Dudley Thought of all of this

Dudley called this phenomenon *the Carrier Nature of Speech*. He meant that the the high bandwidth speech excitation signal is used as a carrier of the slowly changing speech information coded in the spectral envelope. (This is akin to the high frequency carrier that is necessary for radio transmission of a comparatively low frequency audio signal.)

23.3 Machine Analysis by Channel Vocoders

Channel Vocoders perform the analysis of the speech signal using a bank of bandpass filters. The first decision to be made is the type, center frequency, and bandwidths of the filters. The following types of filters are often considered:

Bessel - Bessel filters are often too wide as they have a slow falloff.

Elliptic - Eliptic filters are more selective (faster falloff) but have more ripple in the frequency response which can lead to reverberation.

Chebychev - Chebychev filters have similars pros and cons as the eliptic filters.

Butterworth - Butterworth filters offer a compromise between the frequency selectivity of the Elliptic and Chebychev filters and the flatness of the Bessel filter.

FIR - FIR or Finite Impulse Response filters refer to a digital implementation of a tapped delay line where the output is the sum of weighted delayed versions of the input. This is commonly implemented using a digital signal processor (DSP). It has the advantage that the frequency response is more customizable but the disadvantage that it is difficult or impossible to implement it in discrete analog hardware. [1]

The bandwidth of the filters must be chosen carefully. If the filters are narrow then they will resolve the individual harmonics which then no longer is just the spectral shape. If the filters are wider then they tend to model the general spectral shape. (See Figure 23.3) If they are too wide, however, then the spectral shape will be smoothed too much. Something closer to wide bandwidths are used with pitch detection (which will be discussed below) to resolve the harmonics.

The filters are narrower and denser at lower frequencies, as in humans. For instance, at low frequencies they could be as narrow as 80Hz and at higher frequencies as wide as 300Hz or wider. (See Figure 23.2)

The output of the filters is then rectified and low-pass filtered so that only the spectral shape is retained. The low-pass filtering is ok since we only want to periodically sample the energy in that band. It is common to only sample the rectified signal 40 times a second or so, thus reducing its bandwidth to 20Hz.

The complete system is thus composed of a set of bandpass filters followed by rectifiers (not shown) and low pass filters to determine the spectral envelope. Additional components to handle pitch and voicing (excitation) detection as well will be described later. These different pieces of information are then multiplexed together and sent down the communications channel. (See Figure 23.4)

[1] The other 4 types of filters are in the class IIR or Infinite Impulse Response because the output is a sum of weighted delayed versions of the output and the current input. Although IIR typically needs fewer taps and is thus better for analog implementations, it does not guarantee stability and fewer taps affords less control.

23.3.1 How is this Bandwidth Compression?

Consider a signal which is originally at telephone quality – 4kHz bandwidth. If we could process it using 20 of the above bandpass filters that would only require 20x20Hz = 400Hz for a 10-fold reduction in bandwidth. We still have to have to transmit some representation of the excitation signal, however...

23.4 Excitation Detection and Coding

Now that we have determined and transmitted the spectral shape corresponding to the vocal tract filter, we still have to determine and transmit the type and characteristics of the excitation signal. In most basic vocoders, the type is just voiced or unvoiced (V / UV) while the characteristics are the pitch and amplitude of the excitation.

23.4.1 Dudley's Attempt at Pitch Detection

Homer Dudley (remember, the voder guy) attempted automatic pitch detection by assuming that the fundamental frequency was the pitch and was present and thus we just needed to find it. Since the excitation signal whose pitch we are trying to determine contains not only the fundamental but also many harmonics, the effects of the harmonics must be reduced. Since the harmonics are always at higher frequencies, he postulated that a low-pass slope filter as in Figure 23.7a would attenuate the harmonics so that the fundamental would be the strongest frequency and could thus be more readily detected. (Alternatively variable cut-off filters such as in Figure 23.7b could be used. Figure 23.7c is an example of a digitial low pass filter that could be implemented without multiplies. Its response is shown in Figure 23.7d.) The system worked usually but failed if the fundamental frequency was absent. .

23.4.2 Other Methods

Kuhn used a similar system with a set of low-pass filters with different cut-off frequencies to detect pitch in music.

Additional methods include using an interval histogram to detect the periodicity in the signal and determine its pitch.

Neuron-like methods were also investigated. Figure 23.8 shows the operation of a circuit that was excited with spikes corresponding to peaks in the waveform. Use of a blanking period during which the circuit (neuron) could not fire improved performance.

23.4.3 Spectral Flattening

A technique called *spectral flattening* can be used to detect pitch and also to just generate an excitation signal without explicit pitch detection. Recall that the spectrum of the original waveform consists of the slowly varying spectral envelope modulating the excitation signal which has equal harmonics. (See Figure 23.9a.) If the spectrum of the original signal is low-pass filtered then Figure 23.9b results. Spectral flattening can then be used to modify the levels of the harmonics to make them equal and thus spectrally flat. This signal can then be used to detect pitch.

Alternatively, Schroeder et. al. devised a system where this signal could be used to reconstruct harmonics in the higher frequencies in a vocoder. (See Figure 23.10). The lower frequencies were simply transmitted and then filtered.

23.4.4 Why Excitation Coding and Vocoding are Difficult

Many abnormalities in the speech signal make it difficult to determine the excitation parameters accurately. A few examples are listed here and illustrated in Figure 23.5:

1. *Acoustic Noise in Background* - Acoustic noise in the background can often make the V/UV more difficult as well as making it more difficult to determine the correct pitch of voiced sounds.
2. *Telephone Speech* - Telephone speech is filtered to less than about 3 or 4 kHz and is also subject to echos and noise making it quite difficult even by eyeing spectrograms to determine voicing information.
3. *V/UV Transitions* - Sometimes the transition between voiced and unvoiced is sudden and can thus be difficult to detect immediately. Additionally vowel to nasal transitions can be troublesome.
4. *Pitch Variations* - The pitch of speech varies over time and thus must accurately be tracked.
5. *Diplophonic Speakers* - Sometimes glottal pulses are not exactly the same for each pulse, causing the pitch to appear lower than the actual pitch. (See Figure 23.6)

For these reasons and others it is often easier to detect pitch in low-pass filtered speech. Pitch information is usually more prevalent in the lower frequencies.

23.5 How Good are Channel Vocoders Today?

Uncompressed speech sampled at 8kHz using u-law coding requires 64kbps (kilo-bits-per-second). Channel vocoders are able to typically reduce this to about 10kbps with reasonable quality. Research efforts currently involve coding at many rates between 600bps and 16kbps. As we will see in later lectures, new techniques in linear predictive coding can reduce this to 4.8kbps while retaining high quality. At 600bps there is significan reduction in quality but reasonable intelligibility for some applications can be achieved.

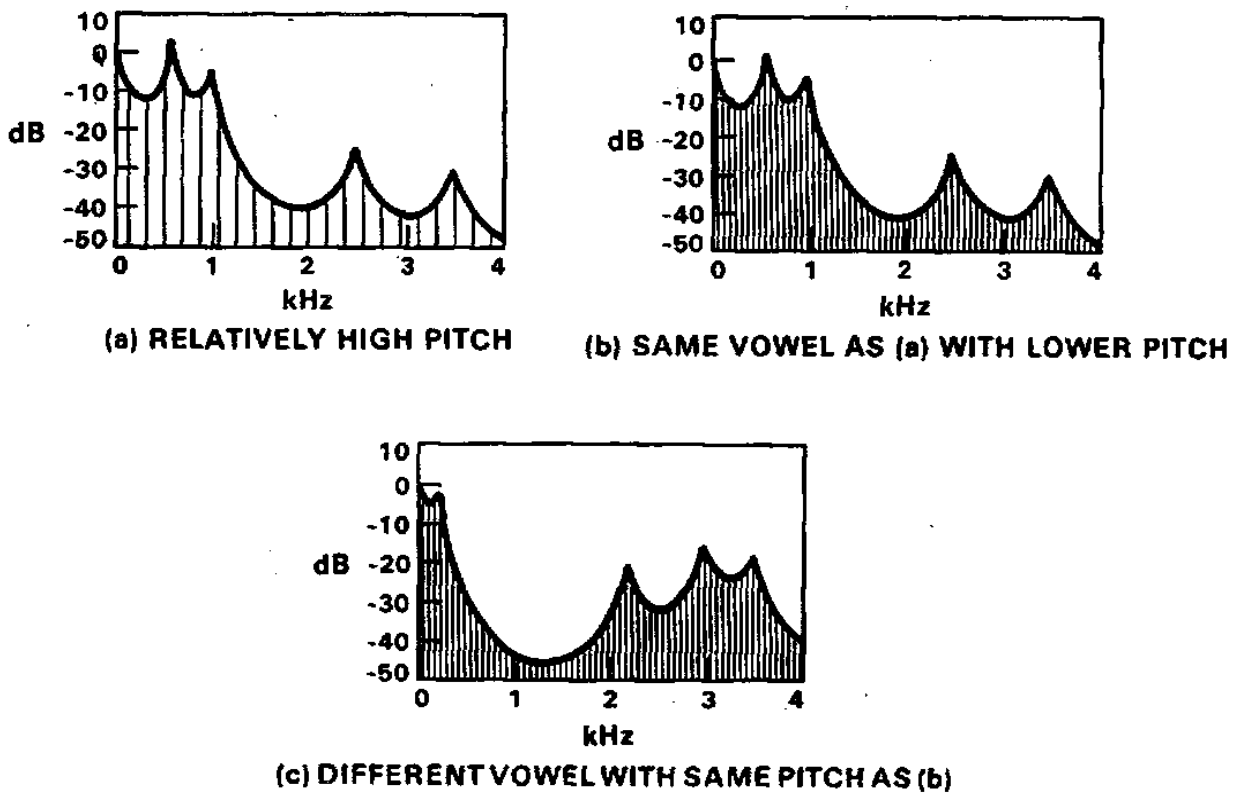

Figure 23.1: Illustrations of fine structure and spectral envelope of vowel sounds.

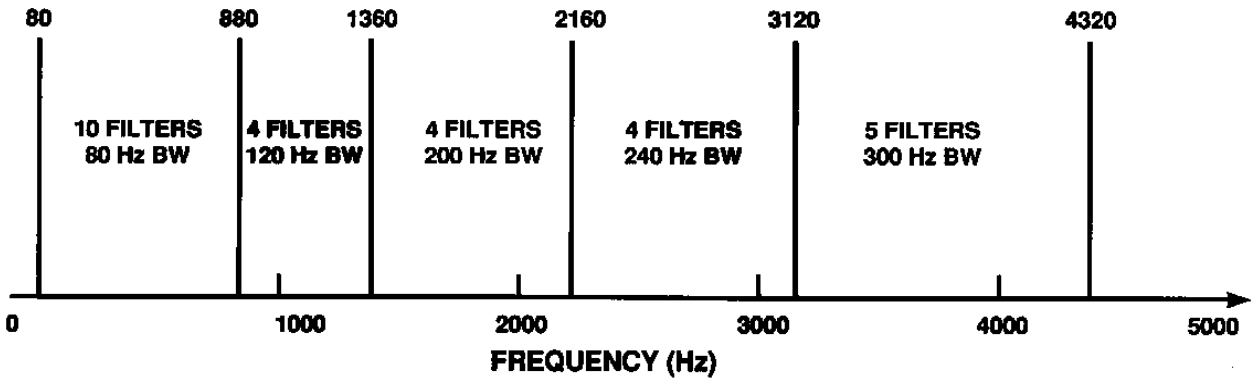

Figure 23.2: Bandpass Filter Spacing and Bandwidths for Vocoder

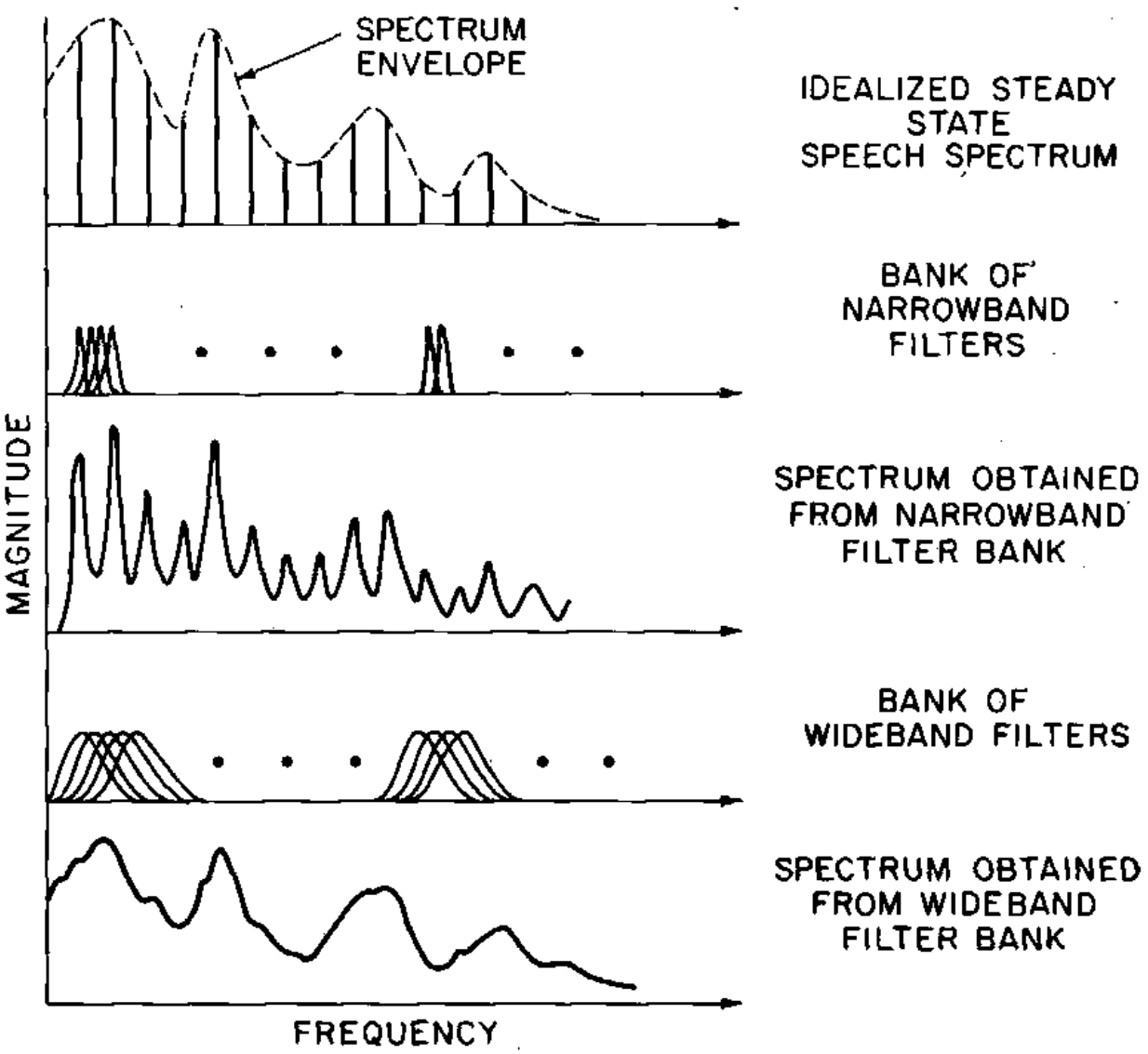

Figure 23.3: Comparison of (Idealized) Measured Spectra for Wide and Narrow Filter Bank Analyzers

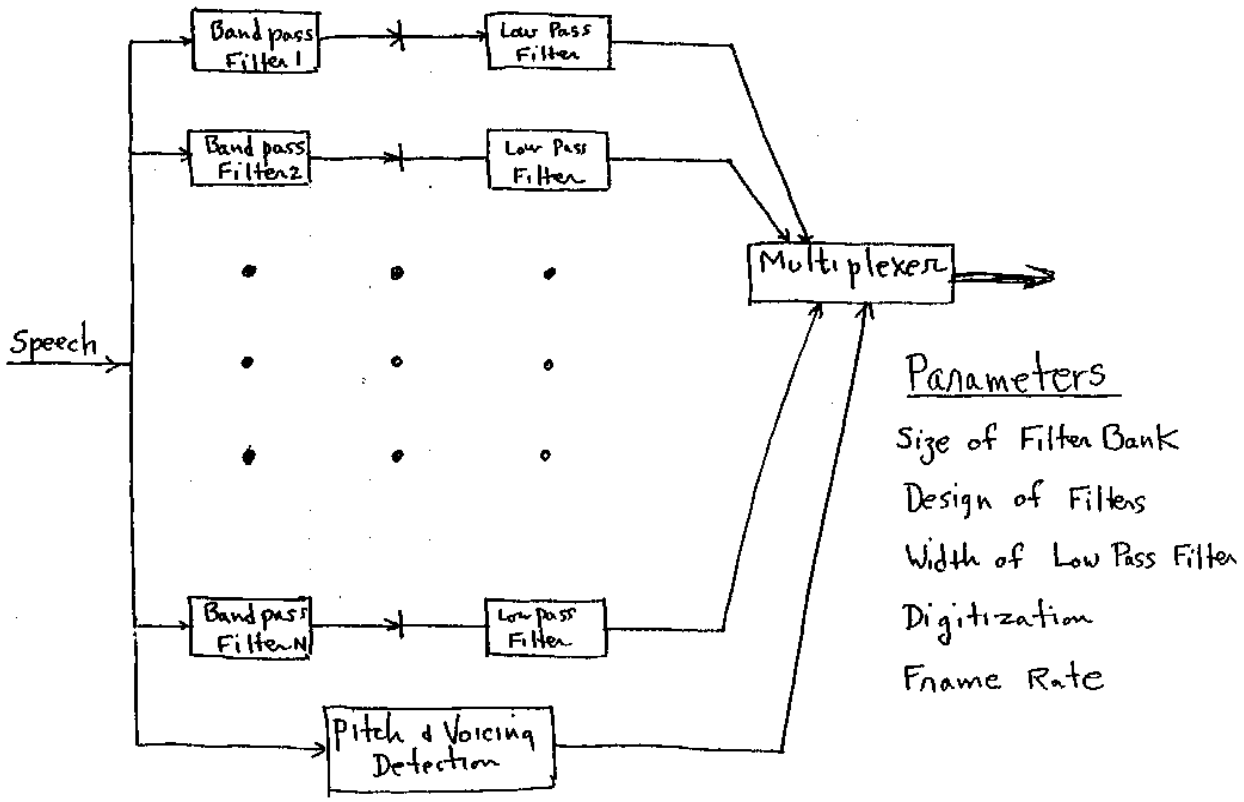

Figure 23.4: Channel Vocoder Analyzer

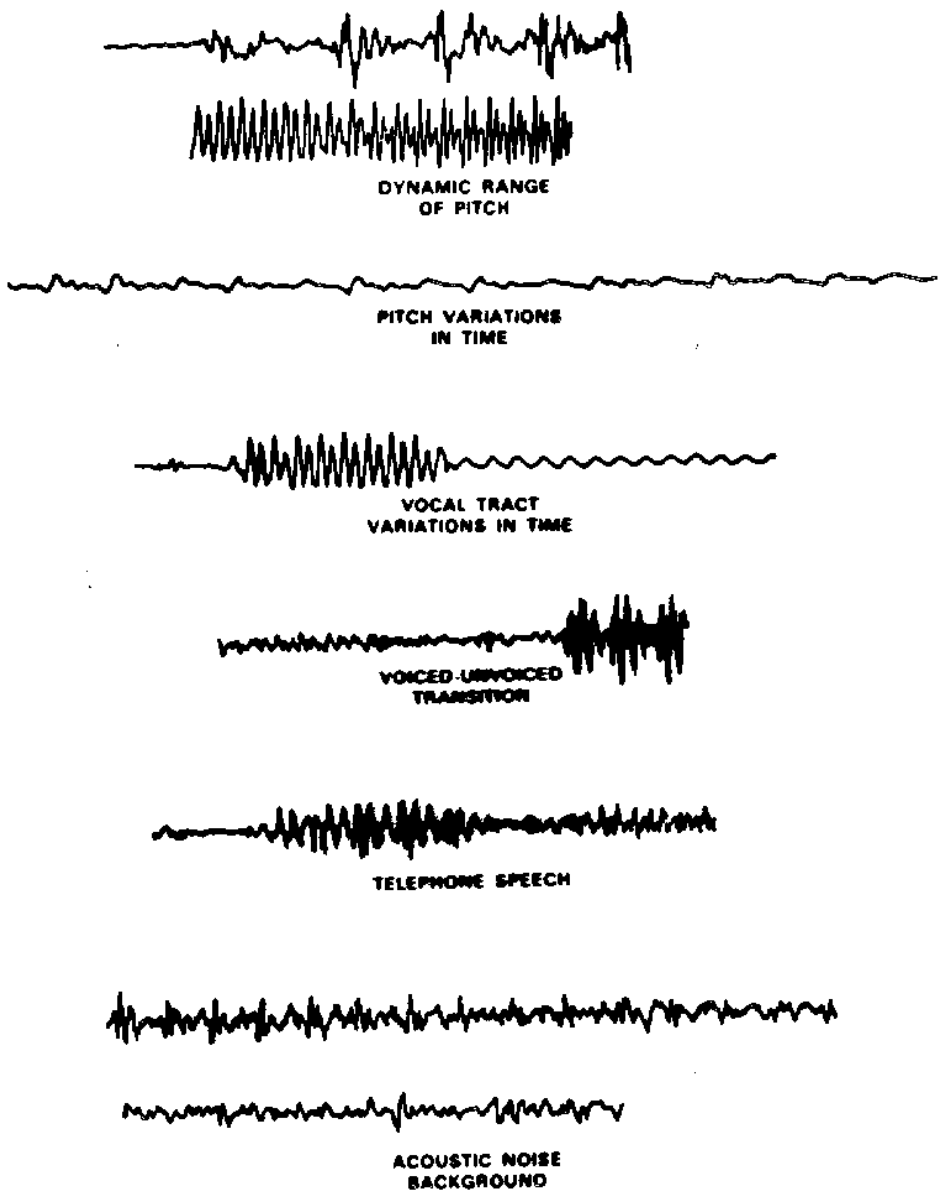

Figure 23.5: Difficult Situations For Pitch Detection

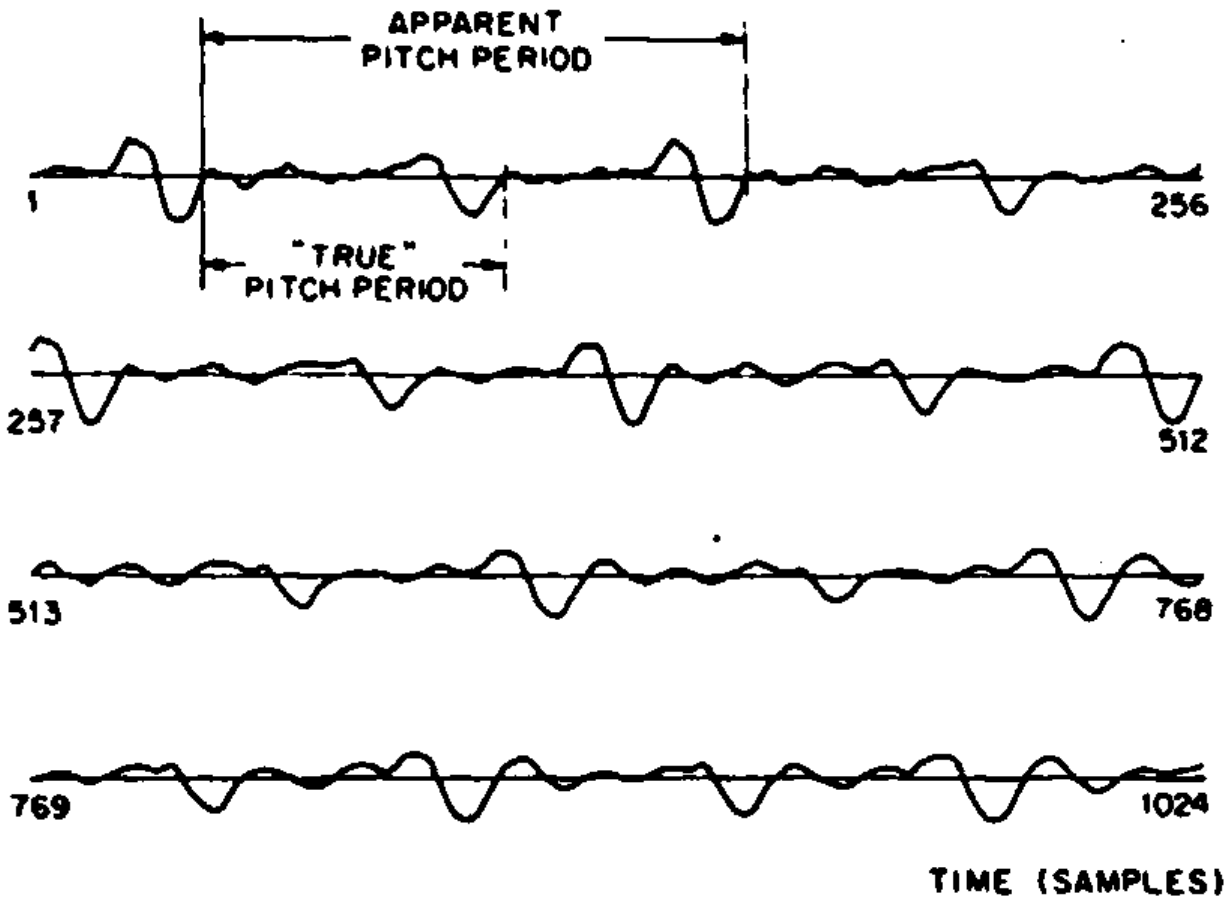

Figure 23.6: Section of the waveform from the diplophonic speaker

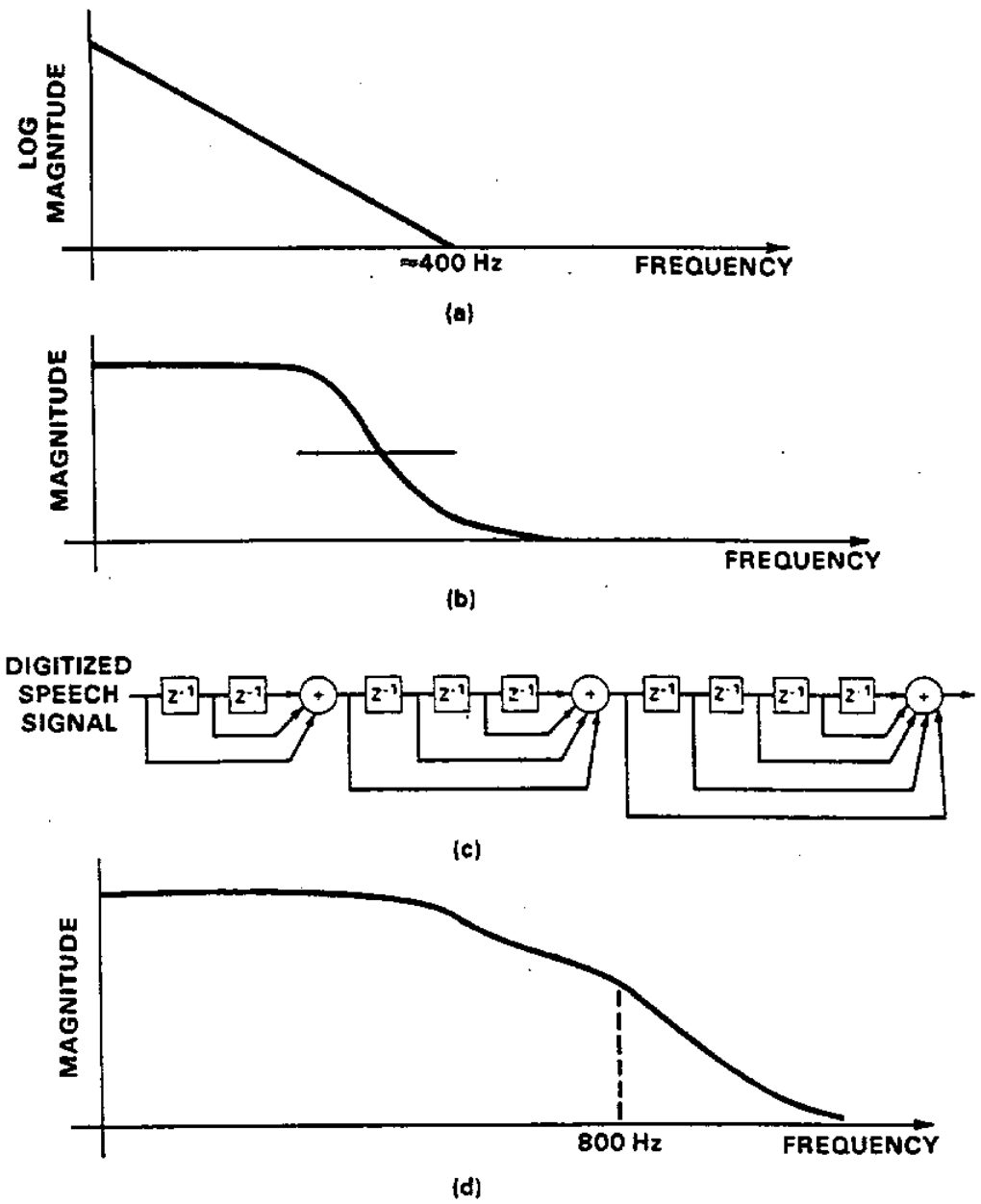

Figure 23.7: Various pitch detection filters: (a) Dudley slope filter. (b) Variable cut-off filter. (c) Frequency response of 3,4,5 filter at a 8-KHz sampling rate

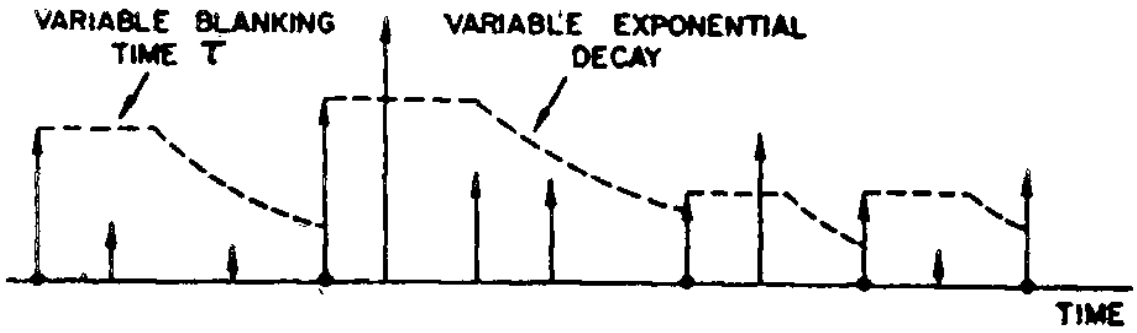

Figure 23.8: Operation of the detection circuit which consists of a variable blanking time during which no pulses are accepted, followed by a variable exponential rundown.

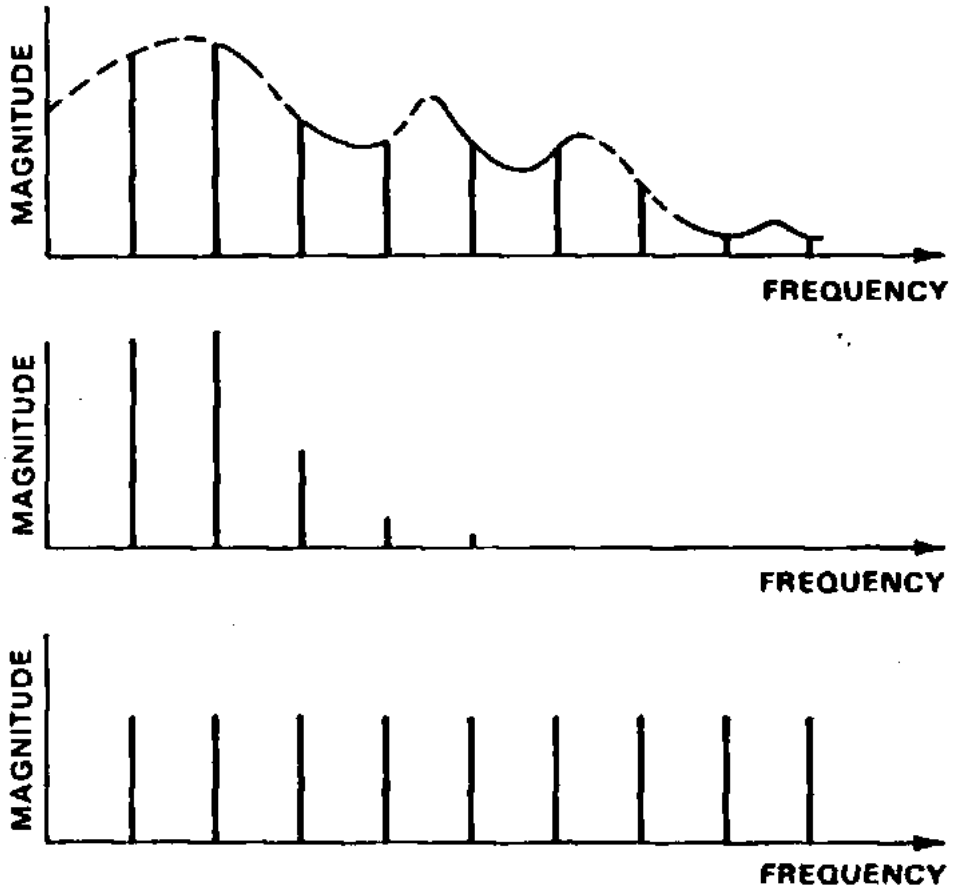

Figure 23.9: Illustration of spectral flattening of the baseband signal to be used as the vocoder excitation: (a) Original Spectrum of periodic waveform, (b) Spectrum after low pass filtering. (c) Spectral flattening.

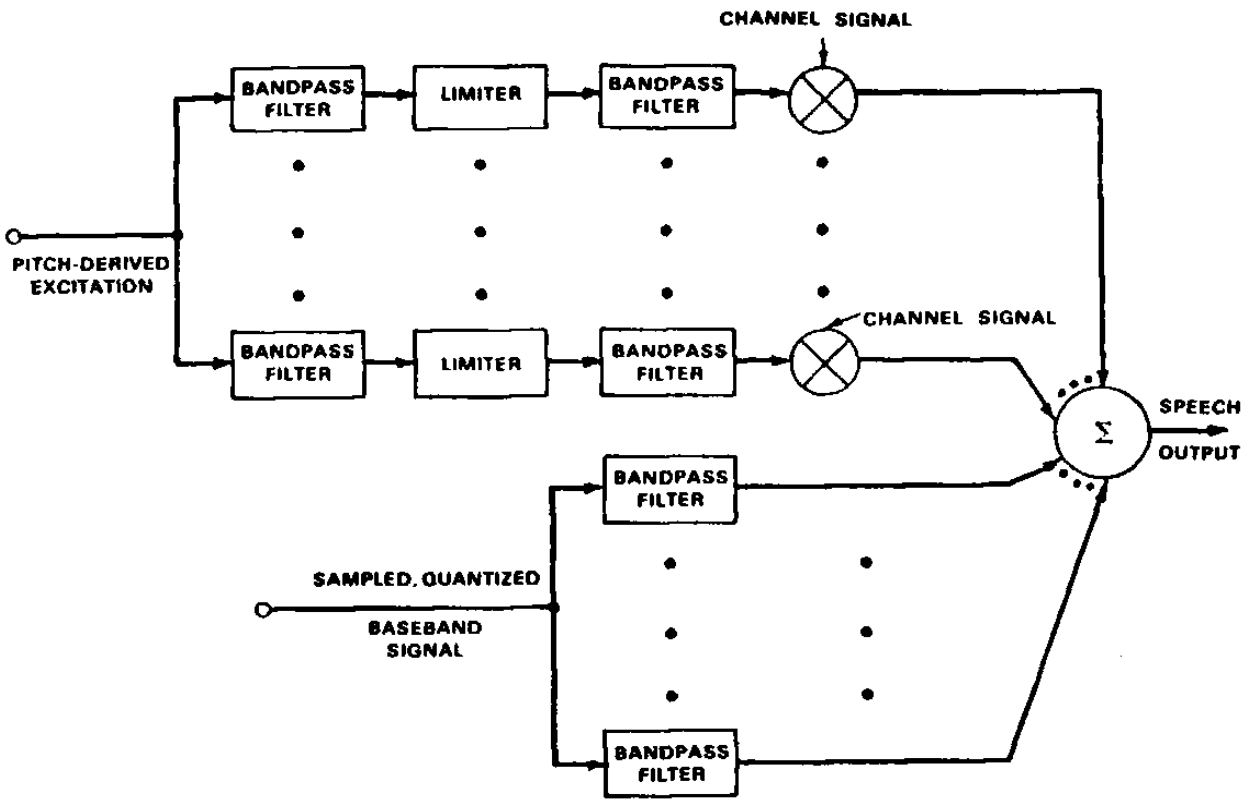

Figure 23.10: Split-band system using a channel vocoder for the higher frequencies