Type: MCQ

Q1. Which of the following is true? (0.5)

1. **The information rate in a typical speech production system increases at each step and that of the speech recognition system decreases at each step
2. The information rate in a typical speech production system decreases at each step and that of the speech recognition system increases at each step
3. The information rate in a typical speech production step and speech recognition step remains the same
4. The information rate in a typical speech production step and speech recognition step cannot be compared

Q2. The utterance of which of the following words has a diphthong in it? (0.5)

1. is
2. car
3. to
4. **buy

Q3. An example of a nasal consonant is (0.5)

1. **m
2. z
3. o
4. p

Q4. Which of the following is false? (0.5)

1. LPC uses frames
2. **LPC uses a series of filters
3. LPC has spectral analysis
4. LPC has parameter conversion

Q5. Which of the following is not an acoustic parameter? (0.5)

1. formant
2. energy in a spectral band
3. duration
4. **distance measure

Q6. TASI is used for (0.5)

1. increase the spectral frequency of each signal
2. decrease the time for transmission of speech signals
3. **Optimize the utilization of multichannel communication media

4. Reduce spectral interference

Q7. Endpoint shifting will result in (0.5)

1. better speech recognition accuracy
2. **lesser speech recognition accuracy
3. increase in computational time
4. decrease in computational time

Q8. Which of the following is not a property of distance measure? (0.5)

1. Positive definiteness
2. **Wavelength
3. symmetry
4. triangle inequality

Q9. Spectral changes that change the perceived sound include (0.5)

1. **changes in formant frequency
2. highpass filtering
3. lowpass filtering
4. notch filtering

Q10. Which of the following is not an unvoiced fricative? (0.5)

1. /f/
2. **/a/
3. /s/
4. /sh/

Type: DES

Q11. Explain the Speech production and Speech recognition system in humans. (4)

Ans.

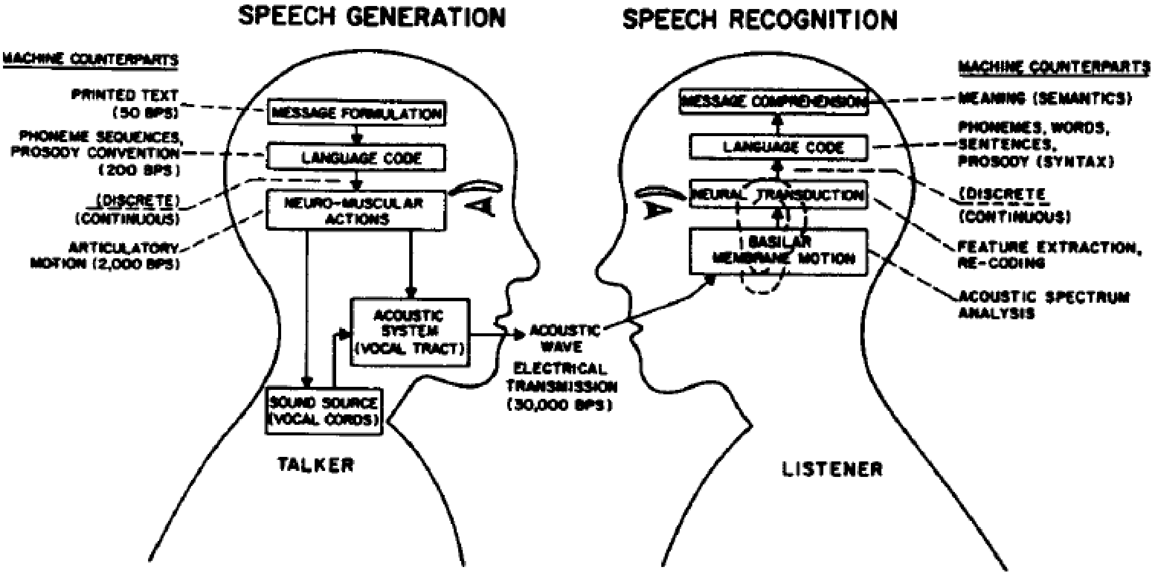


Figure 2.1 Schematic diagram of speech-production/speech-perception process (after Flanagan [unpublished]).

Figure 2.1 shows a schematic diagram of the speech-production/speech-perception process in human beings. The production (speech-generation) process begins when the talker

formulates a message (in his mind) that he wants to transmit to the listener via speech. The machine counterpart to the process of message formulation is the creation of printed text expressing the words of the message. The next step in the process is the conversion of the message into a language code. This roughly corresponds to converting the printed text of the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody markers denoting duration of sounds, loudness of sounds, and pitch accent associated with the sounds. Once the language code is chosen, the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output. The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the lips, jaw, tongue, and velum (a "trapdoor" controlling the acoustic flow to the nasal mechanism).

Once the speech signal is generated and propagated to the listener, the speech-perception (or speech-recognition) process begins. First the listener processes the acoustic signal along the basilar membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process. In a manner that is not well understood, the neural activity along the auditory nerve is converted into a language code at the higher centers of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

Mark: 2 Marks for Figrue, 2 Marks for explanation

Q12. Explain the Neural Networks-based approach to speech recognition. (4)

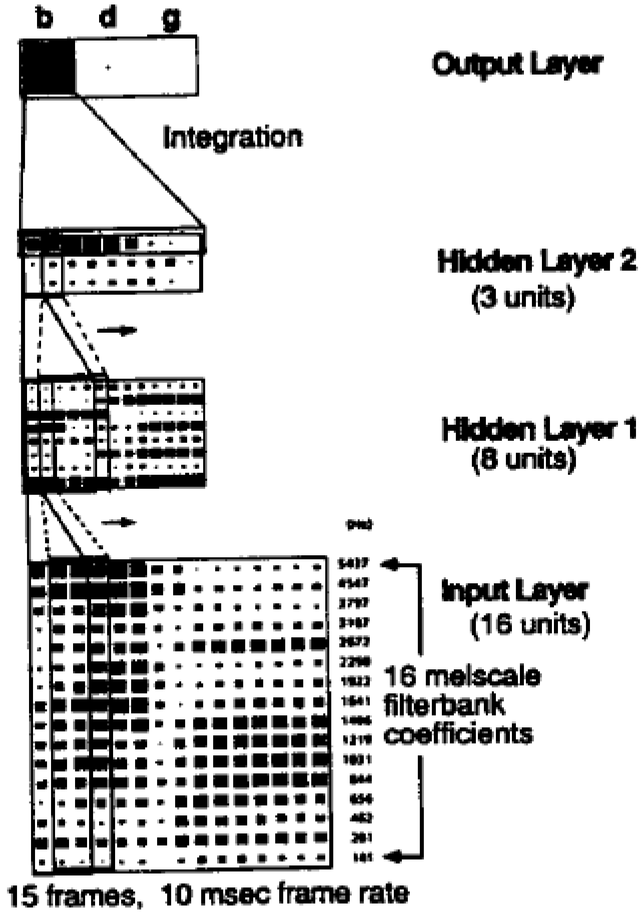


Figure 2.50 A TDNN architecture for recognizing /b/, /d/ and /g/ (after Waibel et al. [14]).

Conventional artificial neural networks are structured to deal with static patterns. As discussed throughout this chapter, speech is inherently dynamic in nature. Hence, some modifications to the simple structures discussed in the previous sections are required for all but the simplest of problems. There is no known correct or proper way of handling speech dynamics within the framework already discussed; however, several reasonable structures have been proposed and studied and we will point out a few such structures in this section.

Perhaps the simplest neural network structure that incorporates speech pattern dynamics is the time delay neural network (TDNN) computation element shown in Figure 2.49 [14]. This structure extends the input to each computational element to include N speech frames (i.e., spectral vectors that cover a duration of $N\Delta$ seconds, where Δ is the time separation between adjacent speech spectra). By expanding the input to N frames (where N is on the order of 15), various types of acoustic-phonetic detectors become practical via succeeding lines in Figure 2.52. Finally, at the end of the sequence, the convolved outputs are summed up and yield a large value, indicating the recognition of the appropriate sound.

Finally, yet a third way of integrating temporal information into a neural network is shown in Figure 2.53. This network is called a hidden control neural network (HCNN) [16] and uses the time varying control, c, as a supplement to the standard input, x, to allow the network properties (input-output relations) to change over time in a well-prescribed manner.

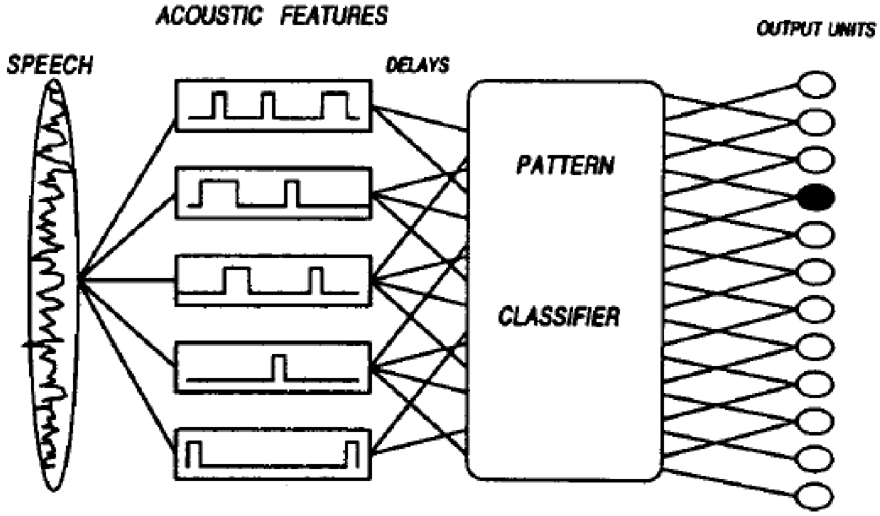


Figure 2.51 A combination neural network and matched filter for speech recognition (after Tank & Hopfield [15]).

succeeding lines in Figure 2.52. Finally, at the end of the sequence, the convolved outputs are summed up and yield a large value, indicating the recognition of the appropriate sound.

Finally, yet a third way of integrating temporal information into a neural network is shown in Figure 2.53. This network is called a hidden control neural network (HCNN) [16] and uses the time varying control, c, as a supplement to the standard input, x, to allow the network properties (input-output relations) to change over time in a well-prescribed manner.

Marks: 1 Marks for each Figure, 2 Marks for Explanation

Q13. Explain the acoustic-phonetic approach to speech recognition. (3)

Ans.

Figure 2.32 shows a block diagram of the acoustic-phonetic approach to speech recognition. The first step in the processing (a step common to all approaches to speech recognition) is the speech analysis system (the so-called feature measurement method), which provides an appropriate (spectral) representation of the characteristics of the time-varying speech signal. The most common techniques of spectral analysis are the class of filter bank methods and the class of linear predictive coding (LPC) methods. The properties of these methods will be discussed in great detail in Chapter 3. Broadly speaking, both of these methods provide spectral descriptions of the speech over time.

The next step in the processing is the feature-detection stage. The idea here is to convert the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. Among the features proposed for recognition are nasality (presence or absence of nasal resonance), frication (presence or absence of random excitation in the speech), formant locations (frequencies of the first three resonances), voiced-unvoiced classification (periodic or aperiodic excitation), and ratios of high- and low-frequency energy. Many proposed features are inherently binary (e.g., nasality, frication, voiced-unvoiced); others are continuous (e.g., formant locations, energy ratios). The feature-detection stage usually consists of a set of detectors that operate in parallel and use appropriate processing and logic to make the decision as to presence or absence, or

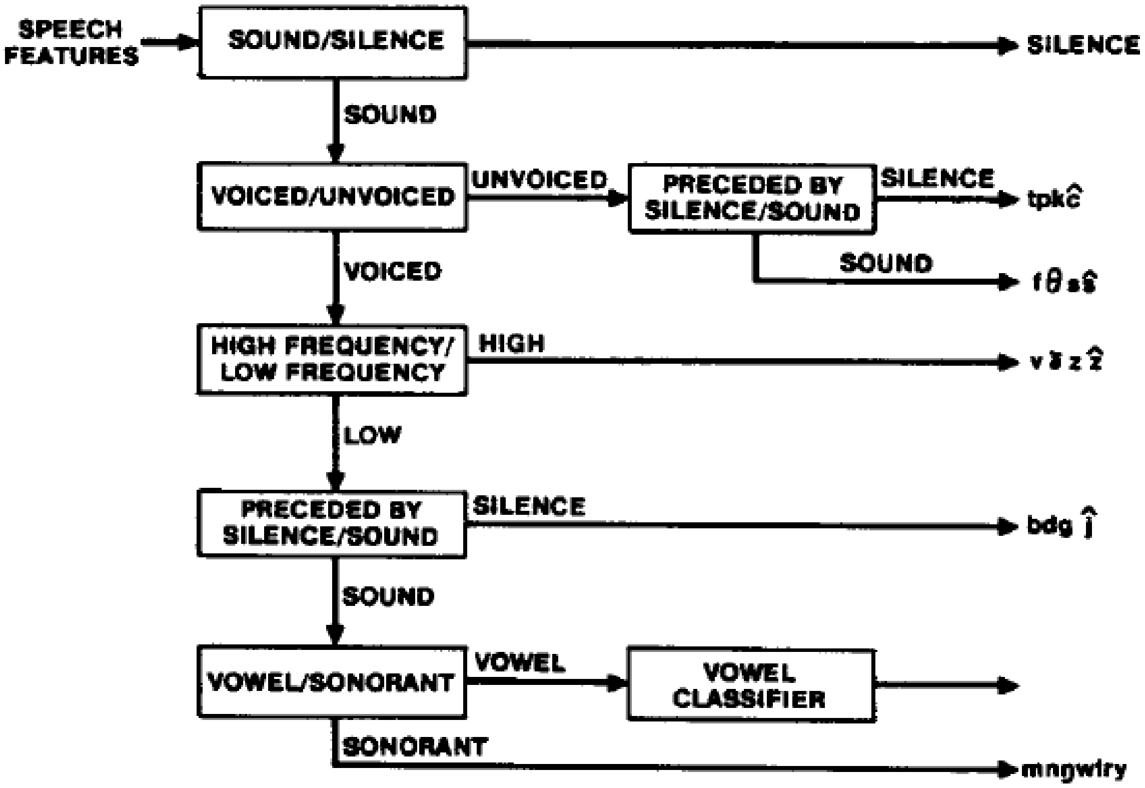


Figure 2.34 Binary tree speech sound classifier.

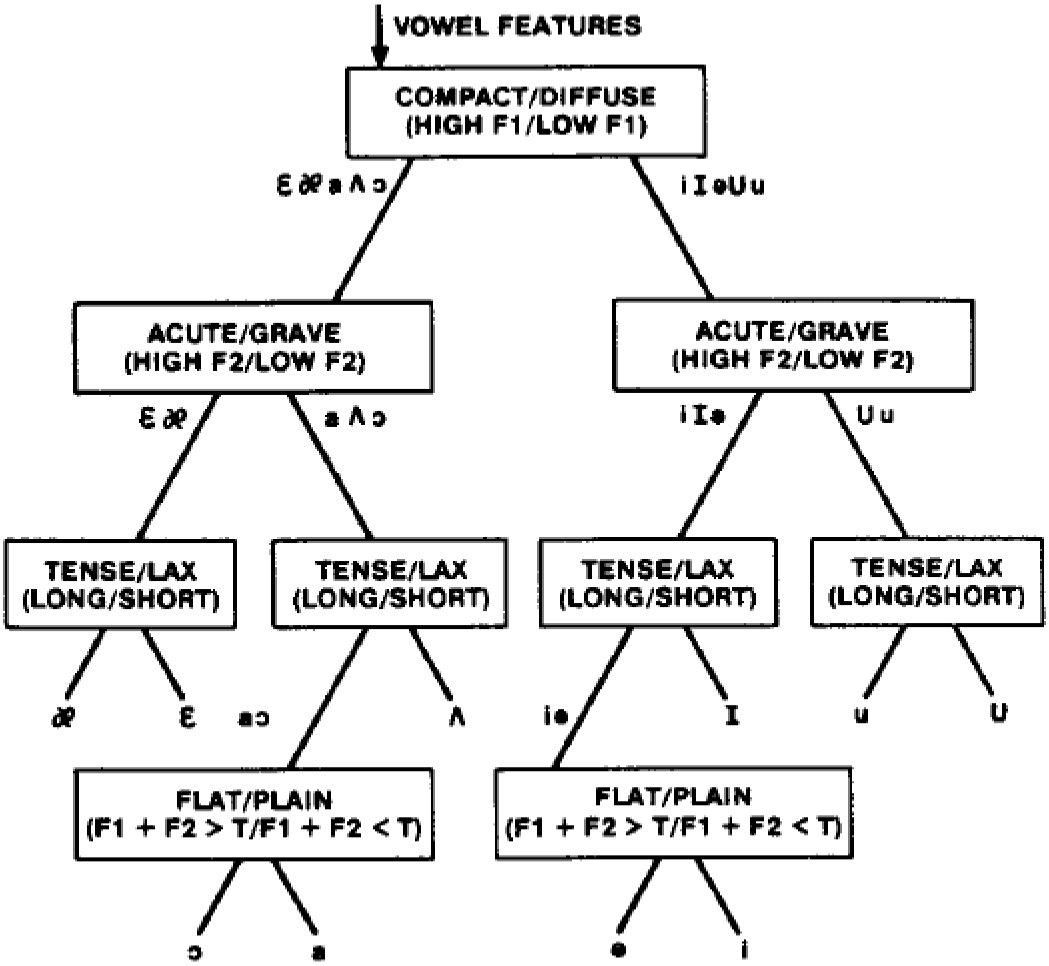
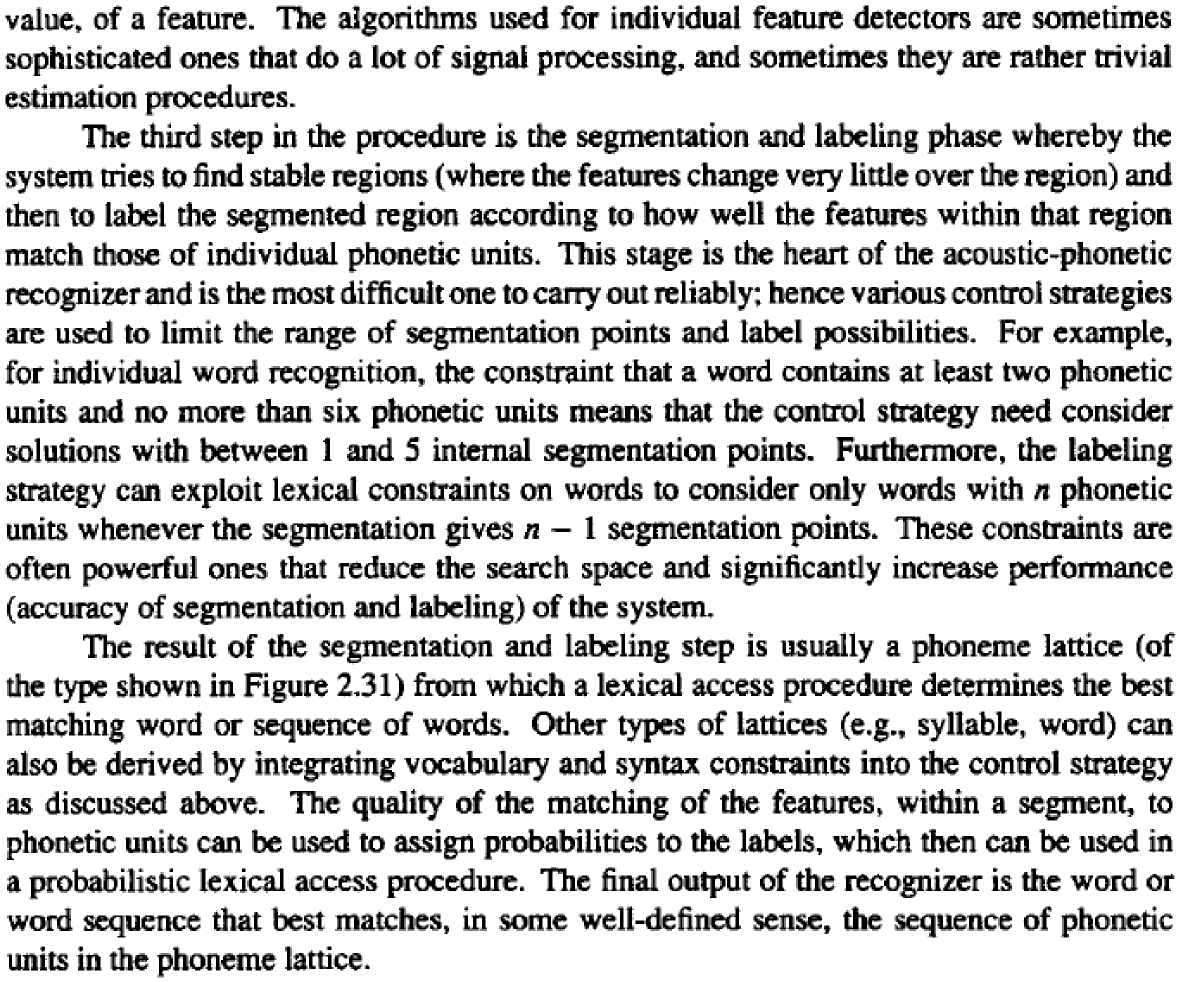


Figure 2.33 Acoustic-phonetic vowel classifier.



value, of a feature. The algorithms used for individual feature detectors are sometimes sophisticated ones that do a lot of signal processing, and sometimes they are rather trivial estimation procedures.

The third step in the procedure is the segmentation and labeling phase whereby the system tries to find stable regions (where the features change very little over the region) and then to label the segmented region according to how well the features within that region match those of individual phonetic units. This stage is the heart of the acoustic-phonetic recognizer and is the most difficult one to carry out reliably; hence various control strategies are used to limit the range of segmentation points and label possibilities. For example, for individual word recognition, the constraint that a word contains at least two phonetic units and no more than six phonetic units means that the control strategy need consider solutions with between 1 and 5 internal segmentation points. Furthermore, the labeling strategy can exploit lexical constraints on words to consider only words with n phonetic units whenever the segmentation gives $n - 1$ segmentation points. These constraints are often powerful ones that reduce the search space and significantly increase performance (accuracy of segmentation and labeling) of the system.

The result of the segmentation and labeling step is usually a phoneme lattice (of the type shown in Figure 2.31) from which a lexical access procedure determines the best matching word or sequence of words. Other types of lattices (e.g., syllable, word) can also be derived by integrating vocabulary and syntax constraints into the control strategy as discussed above. The quality of the matching of the features, within a segment, to phonetic units can be used to assign probabilities to the labels, which then can be used in a probabilistic lexical access procedure. The final output of the recognizer is the word or word sequence that best matches, in some well-defined sense, the sequence of phonetic units in the phoneme lattice.

Marks: 1 Mark for each Figure, 1 Mark for Explanation

Q14. Compare and contrast the Bank of Filters method of spectral analysis with the Linear Predictive Coding method. (3)

Bank of Filters Method	**Linear Predictive Coding**
It uses Filters	It does not use Filters
It does not break input into frames	It breaks the input into frames
It is less efficient	It is better than Bank of Filters in terms of efficiency
It does not use all-pole model constraint	It uses an all-pole model constraint

Marks: Any 3, each one 1 Mark

Q15. Explain the Linear Predictive Coding method of Spectral analysis. (3)

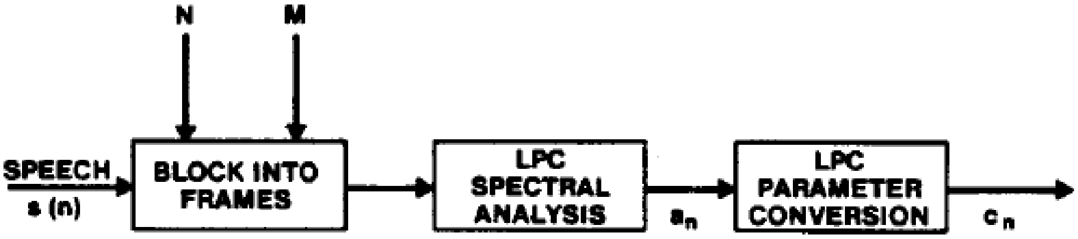
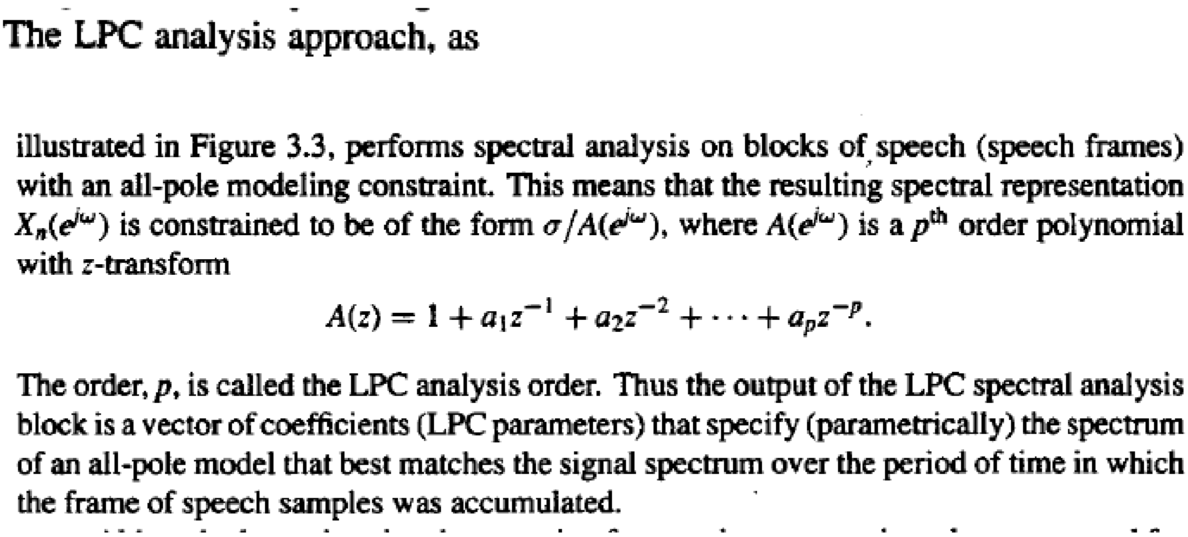


Figure 3.3 LPC analysis model.



The LPC analysis approach, as

illustrated in Figure 3.3, performs spectral analysis on blocks of speech (speech frames) with an all-pole modeling constraint. This means that the resulting spectral representation $X_n(e^{j\omega})$ is constrained to be of the form $\sigma/A(e^{j\omega})$, where $A(e^{j\omega})$ is a p^{th} order polynomial with z-transform

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_p z^{-p}.$$

The order, p, is called the LPC analysis order. Thus the output of the LPC spectral analysis block is a vector of coefficients (LPC parameters) that specify (parametrically) the spectrum of an all-pole model that best matches the signal spectrum over the period of time in which the frame of speech samples was accumulated.

Marks: 1.5 Mark for figure, 1.5 Mark for Explanation

Q16. Explain the factors that make reliable endpoint detection difficult. (3)

with speaking. Unlike the precursor mouth click, the heavy breathing noise is not separated from the speech and therefore makes accurate endpoint detection quite difficult. Figure 4.4 shows an example of a mouth click produced after speaking. Such clicks are often generated inadvertently from the speaker clicking the lips or snapping the tongue after speaking. In all three examples, we see that the energy level of the artifacts is comparable to speech energy levels.

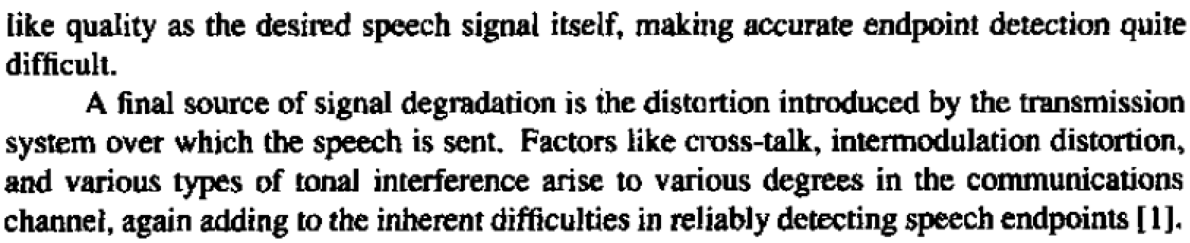
A second factor making reliable speech endpoint detection difficult is the environmental conditions in which the speech is produced. The ideal environment for talking is a quiet room with no acoustic noise or signal generators other than that produced by the speaker. Such an ideal environment is not always practical; hence, one must consider speech produced in noisy backgrounds (as with fans or machinery running), in nonstationary environments (as in the presence of door slams, irregular road noise, car horns), with speech interference (as from TV, radio, or background conversations), and in hostile circumstances (when the speaker is stressed, such as when navigating an airplane or while moving at high speeds). Some of these interfering signals possess as much speech-

with speaking. Unlike the precursor mouth click, the heavy breathing noise is not separated from the speech and therefore makes accurate endpoint detection quite difficult. Figure 4.4 shows an example of a mouth click produced after speaking. Such clicks are often generated inadvertently from the speaker clicking the lips or snapping the tongue after speaking. In all three examples, we see that the energy level of the artifacts is comparable to speech energy levels.

A second factor making reliable speech endpoint detection difficult is the environmental conditions in which the speech is produced. The ideal environment for talking is a quiet room with no acoustic noise or signal generators other than that produced by the speaker. Such an ideal environment is not always practical; hence, one must consider speech produced in noisy backgrounds (as with fans or machinery running), in nonstationary environments (as in the presence of door slams, irregular road noise, car horns), with speech interference (as from TV, radio, or background conversations), and in hostile circumstances (when the speaker is stressed, such as when navigating an airplane or while moving at high speeds). Some of these interfering signals possess as much speech-

For speech produced in the most benign circumstances—that is, carefully articulated and spoken in a relatively noise-free environment—accurate detection of speech is a simple problem. However, this is not usually the case. In practice, one or more problems usually make accurate endpoint detection difficult. One particular class of problems is those attributed to the speaker and to the manner of producing the speech. For example, during articulation, the talker often produces sound artifacts, including lip smacks, heavy breathing, and mouth clicks and pops. Figures 4.2 through 4.4 show examples of this type of humanly produced sound artifact. The top part of each figure is the energy contour of the utterance on a logarithmic (dB) scale, and the lower part is the time waveform of the corresponding utterance.

Figure 4.2 shows a typical mouth click produced by opening the lips (prior to speaking) when the mouth is relatively dry, thereby causing the lips to pop, (i.e., produce a high-frequency,transient sound). Figure 4.3 shows a high level of breath noise produced at the end of speaking, caused by the speaker's heavy breathing. This artifact typically occurs when a speaker is short of breath (usually from exertion) and combines heavy breathing



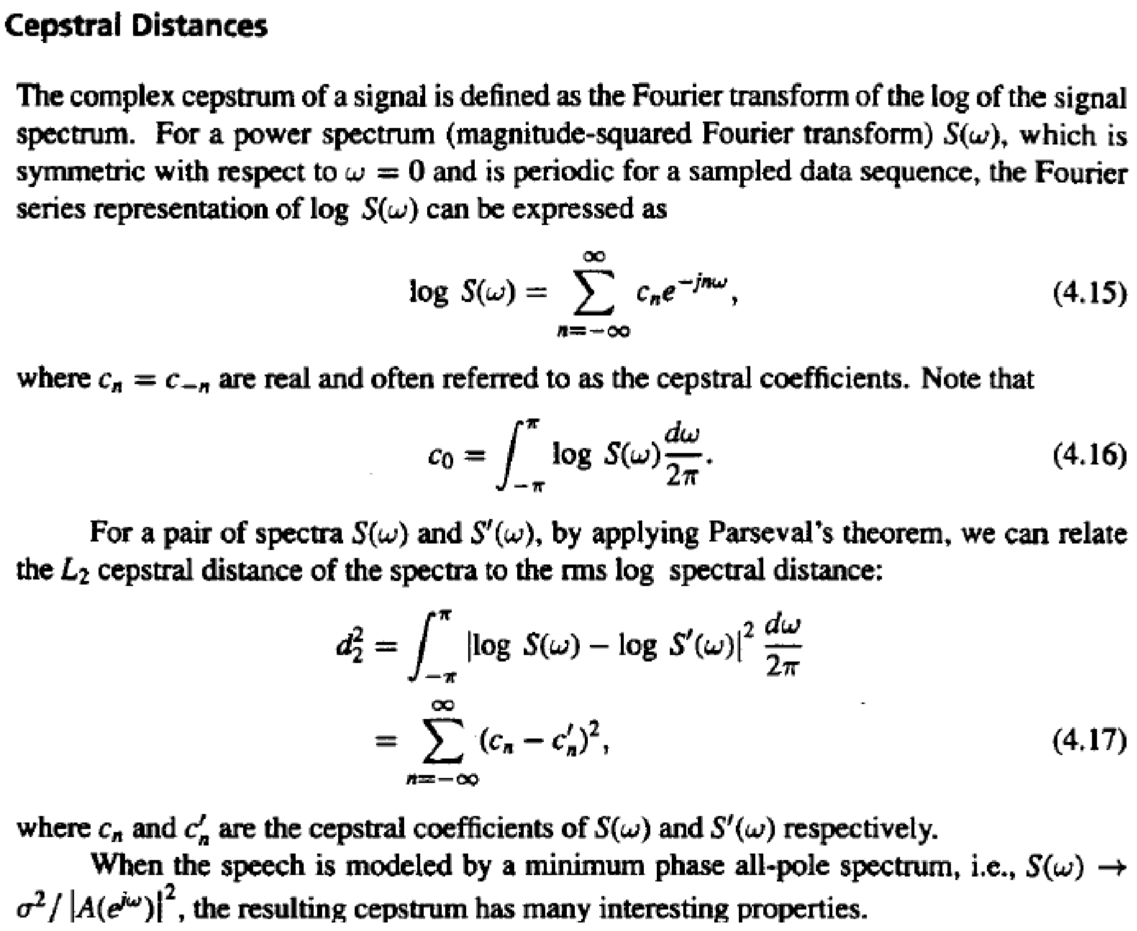
like quality as the desired speech signal itself, making accurate endpoint detection quite difficult.

A final source of signal degradation is the distortion introduced by the transmission system over which the speech is sent. Factors like cross-talk, intermodulation distortion, and various types of tonal interference arise to various degrees in the communications channel, again adding to the inherent difficulties in reliably detecting speech endpoints [1].

Marks: 1 Mark for each point

Q17. Explain Cepstral Distance measure. (3)

Ans.



Cepstral Distances

The complex cepstrum of a signal is defined as the Fourier transform of the log of the signal spectrum. For a power spectrum (magnitude-squared Fourier transform) $S(\omega)$, which is symmetric with respect to $\omega = 0$ and is periodic for a sampled data sequence, the Fourier series representation of log $S(\omega)$ can be expressed as

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega}, \tag{4.15}$$

where $c_n = c_{-n}$ are real and often referred to as the cepstral coefficients. Note that

$$c_0 = \int_{-\pi}^{\pi} \log S(\omega) \frac{d\omega}{2\pi}. \tag{4.16}$$

For a pair of spectra $S(\omega)$ and $S'(\omega)$, by applying Parseval's theorem, we can relate the L_2 cepstral distance of the spectra to the rms log spectral distance:

$$\begin{aligned} d_2^2 &= \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} \\ &= \sum_{n=-\infty}^{\infty} (c_n - c_n')^2, \end{aligned} \tag{4.17}$$

where c_n and c_n' are the cepstral coefficients of $S(\omega)$ and $S'(\omega)$ respectively.

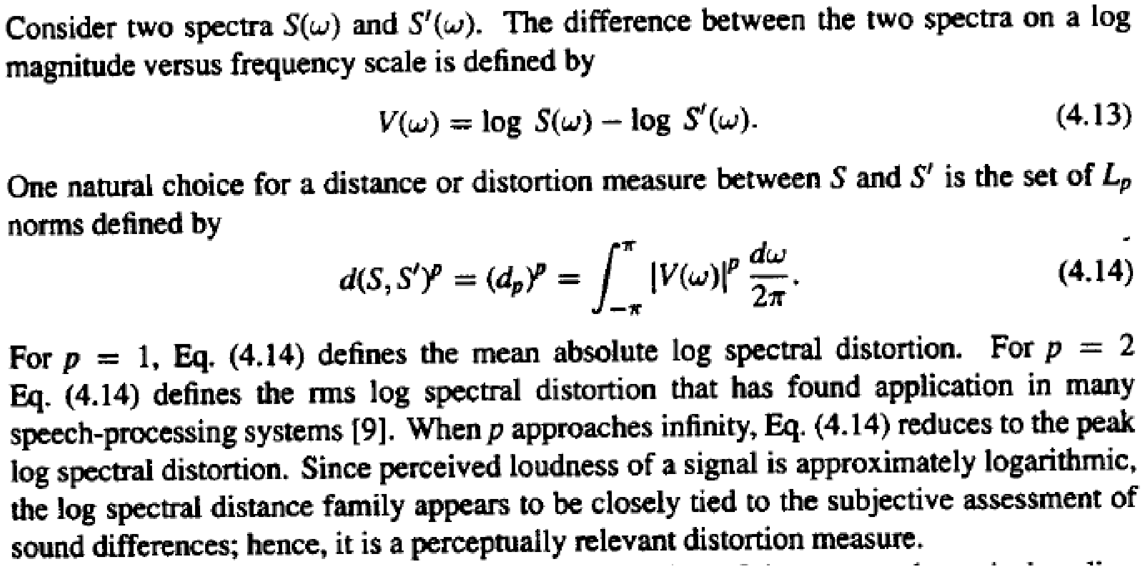
When the speech is modeled by a minimum phase all-pole spectrum, i.e., $S(\omega) \rightarrow \sigma^2 / |A(e^{j\omega})|^2$, the resulting cepstrum has many interesting properties.

Marks: 2 Marks for equation, 1 Mark for Explanation

Q18. Explain Log Spectral Distance measure. (2)

Ans.

Log Spectral Distance



Consider two spectra $S(\omega)$ and $S'(\omega)$. The difference between the two spectra on a log magnitude versus frequency scale is defined by

$$V(\omega) = \log S(\omega) - \log S'(\omega). \tag{4.13}$$

One natural choice for a distance or distortion measure between S and S' is the set of L_p norms defined by

$$d(S,S')^p = (d_p)^p = \int_{-\pi}^{\pi} |V(\omega)|^p \frac{d\omega}{2\pi}. \tag{4.14}$$

For $p = 1$, Eq. (4.14) defines the mean absolute log spectral distortion. For $p = 2$ Eq. (4.14) defines the rms log spectral distortion that has found application in many speech-processing systems [9]. When p approaches infinity, Eq. (4.14) reduces to the peak log spectral distortion. Since perceived loudness of a signal is approximately logarithmic, the log spectral distance family appears to be closely tied to the subjective assessment of sound differences; hence, it is a perceptually relevant distortion measure.

Marks: 1 Mark for Equation, 1 Mark for explanation