13

SPEECH PROCESSING

S peech sounds are sensations of air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the resonance of the vocal tract as the air is pushed out through the lips and nose.

Speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration) to convey information about words, speaker identity, accent, expression, style of speech, emotion and the state of health of the speaker. All this information is conveyed primarily within the traditional telephone bandwidth of 4 kHz. The speech energy above 4 kHz mostly conveys audio quality and sensation.

In this chapter the fundamentals of speech signals and speech production and perception are studied. We study the mechanisms that produce and convey phonetic speech sounds and examine the acoustic correlates of speaker characteristics such as gender, accent and emotion. The spectral and temporal structures of speech are studied and the most commonly used models and features for capturing speech characteristics in time and frequency are introduced. Speech coding methods for improving bandwidth utilisation and power efficiency in mobile communication are covered. Finally, we study automatic speech recognition for a simple voice-dialling application.

13.1 Speech Communication

Speech is the most natural form of human communication. Speech is one of the most information-laid signals; speech sounds have a rich and multi-layered temporal-spectral variation that convey words, intention, expression, intonation, accent, speaker identity, gender, age, style of speaking, state of health of the speaker and emotion.

Speech sounds are produced by air pressure vibrations generated by pushing inhaled air from the lungs through the vibrating vocal cords and vocal tract and out from the lips and nose airways. The air is modulated and shaped by the vibrations of the glottal cords, the resonance of the vocal tract and nasal cavities, the position of the tongue and the openings and closings of the mouth.

Just as the written form of a language is a sequence of elementary alphabet, speech is also a sequence of elementary acoustic sounds or symbols known as phonemes that convey the spoken form of a language. There are about 40-60 phonemes in the English language from which a very large number of spoken words can be constructed. Note that in practice the production of each phonemic sound is affected by the context of the neighbouring phonemes.

Speech signals convey much more than spoken words. The information conveyed by speech is multi-layered and includes time-frequency modulation of such carriers of information as formants and pitch intonation. Formants are the resonances of vocal tract and pitch is the sensation of the fundamental frequency of the opening and closings of the glottal folds. The information conveyed in speech includes the followings:

(a) Acoustic phonetic symbols. These are the most elementary speech units from which larger speech units such as syllables and words are formed. Some words have only two phones such as 'me', 'you','he'.

(b) Prosody. These are rhythms of speech mostly intonation signals carried by changes in the pitch trajectory and stress. Prosody help to signal such information as the boundaries between segments of speech, link sub-phrases and clarify intention and remove ambiguities such as whether a spoken sentence is a statement or a question.

(c) Gender information. Gender is conveyed by the pitch (related to the fundamental frequency of voiced sounds) and the size and physical

characteristics of the vocal tract. Due to differences in vocal anatomy, female voice has higher resonance frequencies and a higher pitch.

(d) Age, conveyed by the effects of the size and the elasticity of the vocal cords and vocal tract, and the pitch. The pitch of voice o children can be more than 300 Hz.

(e) Accent, broadly conveyed through: (i) changes in the pronunciation dictionary in the form of substitution, deletion or insertion of phoneme units in the "standard" transcription of words (e.g. Australian *todie* pronunciation of *today* or US ***Jaan*** pronunciation of ***John***) and (ii) systematic changes in speech resonance frequencies (formants), pitch intonation, duration, emphasis and stress.

(f) Speaker's identity conveyed by the physical characteristics of a person's vocal folds, vocal tract, pitch intonations and stylistics.

(g) Emotion and health, conveyed by changes in: vibrations of vocal fold, vocal tract resonance, duration and stress and by the dynamics of pitch and vocal tract spectrum.

In the remainder of this chapter we will study how various acoustic correlates of speech and speaker can be modelled and used for speech processing applications.

13.2 Acoustic Theory of Speech: The Source-Filter Model

An outline of the anatomy of the human speech production system is shown in Fig. (13.1). It consists of the lungs, larynx, vocal tract cavity, nasal cavity, teeth, lips, and the connecting tubes. The combined voice production mechanism produces the variety of vibrations and spectral-temporal compositions that form different speech sounds.

The act of production of speech begins with exhaling (inhaled) air from the lung. Without the subsequent modulations, this air will sound like a random noise with no information. The information is first modulated onto the passing air by the manner and the frequency of closing and opening of the glottal folds. The output of the glottal fold is the excitation signal to the vocal tract which is further shaped by the resonances of the vocal tract and the effects of the nasal cavities and the teeth and lips.

The vocal tract is bounded by hard and soft tissue structures. These structures are either essentially immobile, such as the hard palate and teeth,
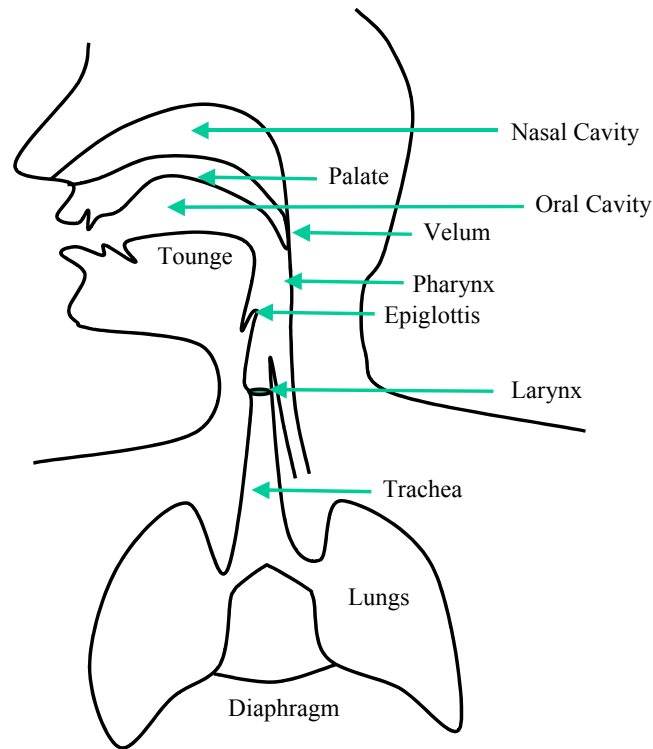
Figure 13.1 Illustration of anatomy of speech production.

or movable. The movable structures associated with speech production are also called *articulators*. The tongue, lips, jaw, and velum are the primary articulators; movement of these articulators appear to account for most of the variations in the vocal tract shape associated with speaking. However, additional structures are capable of motion as well. For instance, the glottis can be moved up or down to shorten or lengthen the vocal tract and hence change its frequency response.

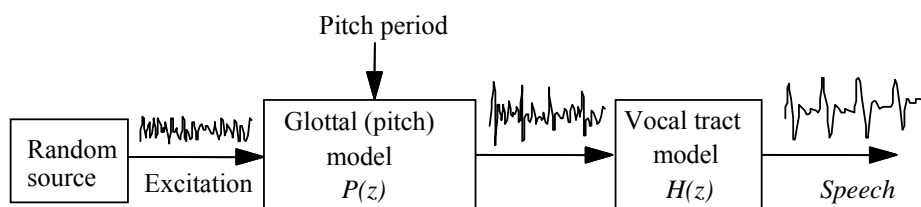Speech sounds result from a combination of a source of sound energy

Figure 13.2 A source–filter model of speech production.

(the larynx) modulated by a time-varying transfer function filter (vocal articulators) determined by the shape and size of the vocal tract. This results in a shaped spectrum with broadband energy peaks. This model is known as the source-filter model of speech production shown in Figure 13.2. In this model the source of acoustic energy is at the larynx, and the vocal tract serves as a time-varying filter whose shape determines the phonetic content of the sounds.

13.2.1 The Source Model

The source signal of speech is the noise-like air from the lungs which is temporally and spectrally shaped by the manner and the frequency of the openings and closings of the glottal folds. There are two broad types of speech sounds as shown in Figure 13.3: *voiced* sounds like an "*e*" pronounced as "*iy*", and *unvoiced* sounds like "*s*".

Voiced sounds are produced by a repeating sequence of opening and closing of glottal folds with a frequency of between 40 (e.g. for a low frequency gravel male voice) to 600 (e.g. for female children's voice) cycles per second (Hz) depending on the speaker, the phoneme and the linguistic and emotional/expressional context. First, as the air is pushed out from the
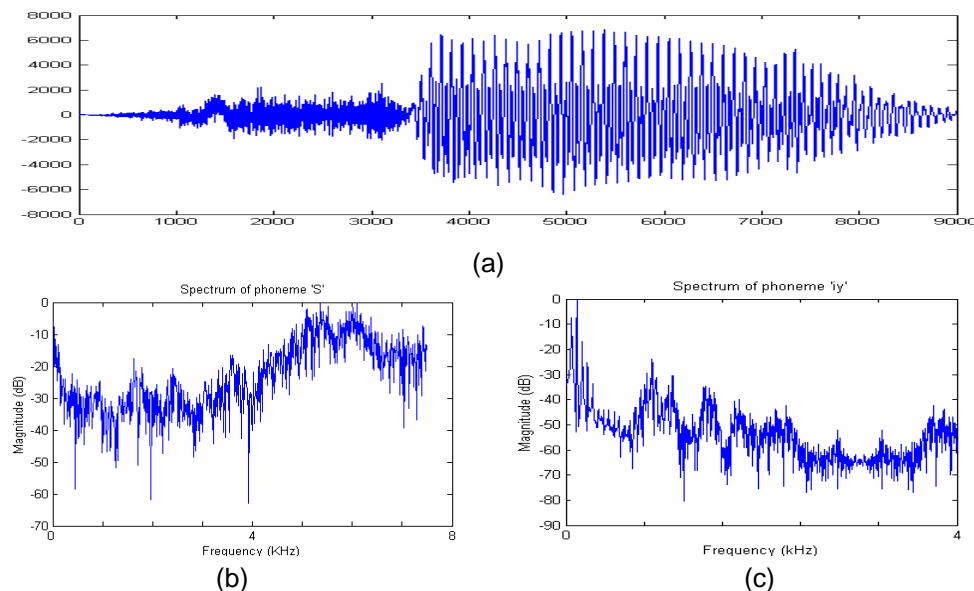

(a)

(b)   (c)

Figure 13.3 (a) Acoustic production of the word **sea** (pronounced *s-iy*), (b) spectrum of the unvoiced segment "*s*", and (c) spectrum of the voiced speech segment "*iy*".

lungs the vocal cords are brought together, temporarily blocking the airflow from the lungs and leading to increased sub-glottal pressure. When the sub-glottal pressure becomes greater than the resistance offered by the vocal folds, the folds open and let out a pulse of air. The folds then close rapidly due to a combination of factors, including their elasticity, laryngeal muscle tension, and the Bernoulli effect of the air stream. If the process is maintained by a steady supply of pressurized air, the vocal cords will continue to open and close in a quasi-periodic fashion. As they open and close, the pulses of air flow through the glottal opening as shown in Figure 13.4.

The periodicity of the glottal pulses determines the fundamental frequency ($F0$) of the laryngeal source and contributes to the perceived pitch of the sound. The time-variations of glottal pulse period convey the expressional content, the intonation, the stress and emphasis in speech signals. In normal speech the fundamental frequency (pitch) changes constantly, providing linguistic and speaker information, as in the different intonation patterns associated with questions and statements, or information about the emotional content, such as differences in speaker mood e.g. calmness, excitement, sadness etc.

Figure (13.3) shows an example of a speech segment containing an unvoiced sound "*s*" and a voiced sound "*iy*". Note that the spectrum of voiced sounds is shaped by the resonance of the vocal tract filter and contains the harmonics of the quasi-periodic glottal excitation, and has most of its power in the lower frequency bands, whereas the spectrum of unvoiced sounds is non-harmonic and usually has more energy in higher frequency bands. The shape of the spectrum of the input to vocal tract filter is determined by the details of the opening and closing movements of the vocal cords, and by the fundamental frequency of the glottal pulses.

For unvoiced sounds (such as consonants) air is passed through some obstacle in the mouth, or is let out with a sudden burst. The position where the obstacle is created depends on which speech sound (i.e. phoneme) is produced. During transitions, and for some mixed-excitation phonemes, the same air stream is used twice: first to make a low-frequency hum with the vocal cords, then to make a high-frequency, noisy hiss in the mouth.

13.2.1.1 Glottal Pulse Model for Voiced Signals

The shape of the glottal pulses of air flow is determined by the manner and the duration of the opening and closing of the glottal folds in each cycle of voice sounds and contributes to the perception of the voice quality and the speaker's identify. The quality of voice and its classification into such types as normal (modal), creaky, breathy, husky, tense etc. depends on the glottal pulse shape.

Figure (13.5) shows the Liljencrants-Fant (LF) model of a glottal pulse and its derivative. The glottal pulse consists of an open phase and a closed
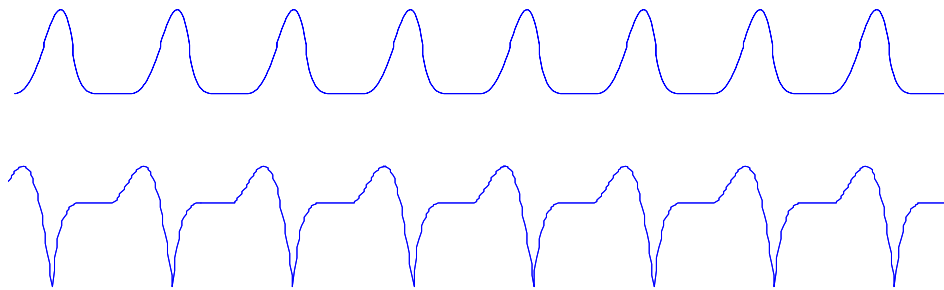
Figure 13.4 – A sequence of glottal pulses of air flow (top and their derivative (bottom).
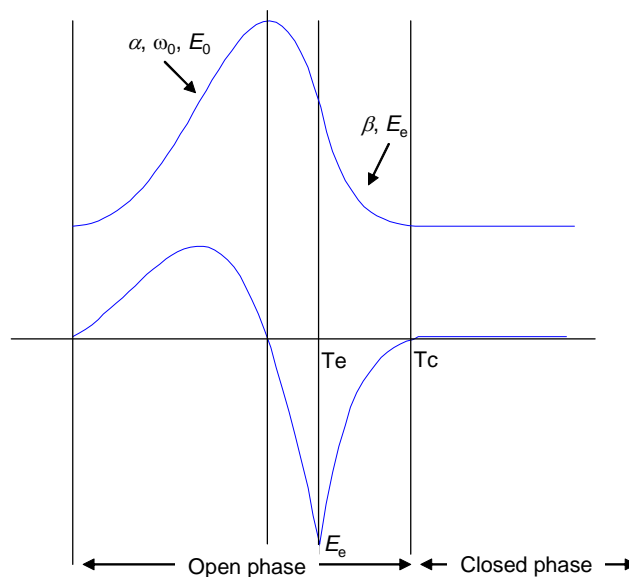
Figure 13.5 – The LF model of a glottal pulse and its derivative.

phase during a pulse or puff of is let through. The open phase of the cycle itself is composed of an opening phase which culminates into the maximum opening of the glottal folds and a closing phase. The maximum negative value of the derivative of the pulse is reached at the point of fastest rate of closing of the glottal folds. The LF model of the derivative of the glottal pulse is defined as

$$v_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin \omega t & 0 \leq t < T_e \\ E_1 \left(e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)} \right) & T_e \leq t < T_c \\ 0 & T_c \leq t \leq T_0 \end{cases} \qquad (13.1)$$

where a segment of less than ¾ of a period of a sine wave, with a frequency of ω and an exponential envelop $E_0 e^{\alpha t}$, is used to model the derivative of the glottal pulse up to the instance T_e where the derivative of the pulse reaches the most negative value which corresponds to the fastest rate of change of the closing of the glottal folds. The final part of the closing phase of the glottal folds, the so called return phase, is modelled by an exponentially decaying function in the second line of Equation (13.1). In Equation (13.1) T_0 is the period of the glottal waveform; $F_0 = 1/T_0$ is the fundamental frequency (pitch) of speech harmonics, and T_c is the instance of closing of the glottal fold. The parameters E_0 and E_1 can be described in terms of the most negative value of the pulse E_e at the instant T_e; $E_o = E_e / e^{\alpha T_e} \sin \omega T_e$ and $E_1 = E_e / [1 - e^{-\beta(T_c - T_e)}]$. The modelling and estimation of the glottal pulse is one of the ongoing challenges of speech processing research.

MatLab Program `function` `GlottalLF()`

13.2.2 The Filter Model

Whereas the source model of speech signals captures and explains the detailed fine structure of speech spectrum, the filter model captures and explains the envelope of the speech spectrum. The reflective and resonance characteristic of the physical space, such as vocal tract, through which a sound wave propagates changes the spectrum of sound and it is perception.

The vocal tract space composed of the oral and the nasal cavities and airways can be viewed as a time-varying acoustic filter that amplifies and filters the sound energy and shapes its frequency spectrum. The resonance frequencies of the vocal tract are called *formants*. The labels of the
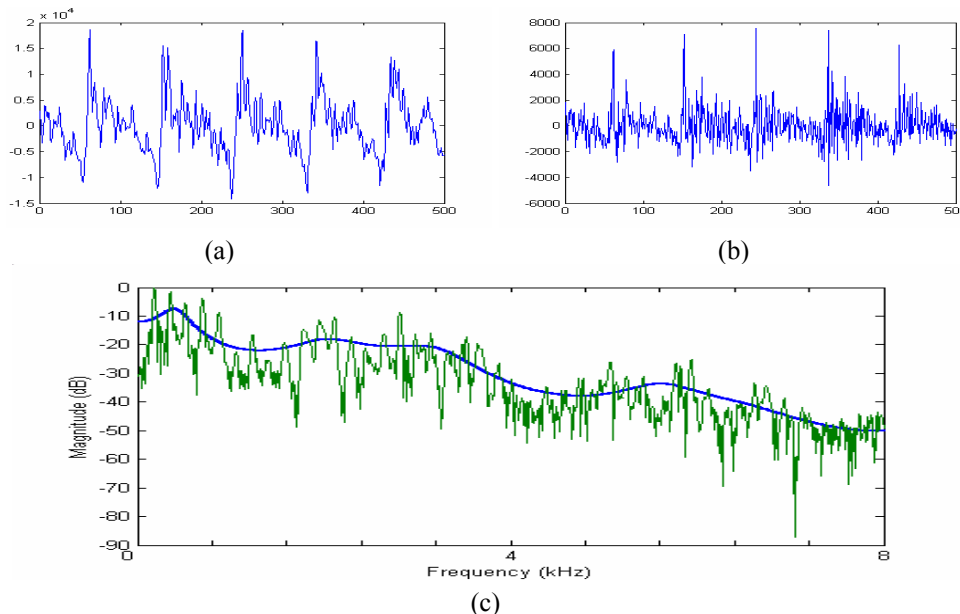
(a)   (b)

(c)

Figure 13.6 (a) a segment of the vowel 'ay', (b) its glottal excitation, and (c) its magnitude Fourier transform and the frequency response of a linear prediction model of the vocal tract.

phonemes are conveyed by the resonance frequencies at formants. Depending on the phoneme sound and the speaker characteristics there are about 3 to 5 formants in voiced sounds.

Formants are dependent on the phonemes but are also affected the overall shape, length, volume and reverberation characteristics of the vocal space and the vocal tract tissues and associated parts i.e nasal cavities, tongues, teeth and lips. The detailed shape of the filter transfer function is determined by the entire vocal tract serving as an acoustically resonant system combined with losses including those due to radiations at the lips.
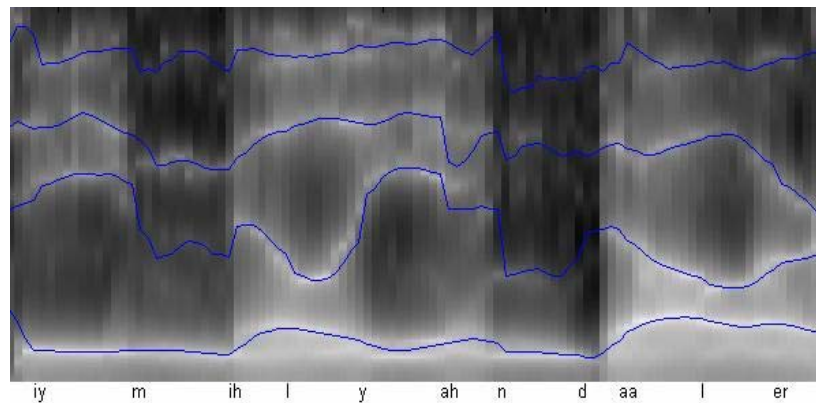

Figure 13.7 An example of formant tracks superimposed on LP spectrogram.

Figure (13.6) shows a segment of a speech signal and its frequency spectrum for the vowel '*ay*'. The formants, which correspond to those frequencies at which the peaks of the frequency response occur, represent the centre points of the main bands of energy that are passed by a particular shape of the vocal tract. In this case they are 500, 2000 and 3000 Hz with bandwidths of 60 to 100 Hz. Figure 13.7 shows formant tracks superimposed on a spectrogram of spectral envelops obtained from the frequency response of linear prediction models of speech. The flexibility of the human vocal tract, in which the articulators can easily adjust to form a variety of shapes, results in the potential to produce a wide range of sounds. For example, the particular vowel quality of a sound is determined mainly by the shape of the vocal tract, and is reflected in the filter impulse response or frequency response function.

13.3 Speech Models and Features

Speech models and features are used for speech coding, recognition, synthesis, and enhancement. For speech coding most commercially used systems are based on linear prediction models. For text to speech synthesis often a harmonic plus noise model of speech used. For speech recognition the most popular features, derived from spectral envelope, are the variants of cepstral features (including FFT-cepstra and LPC-cepstra) and their dynamics in time. The purpose of speech modelling and parameterisation is two fold:

(a) To represent speech in terms of a compact set of parameters for speech coding, recognition, synthesis, or enhancement.

(b) To separate speech parameters, such as spectral envelope and pitch intonation curve, which perform different functions or convey different signals such as accent, emotion, gender etc.

Speech features and parameters can be listed as follows:

(a) Spectral envelope of speech is modelled by the frequency response of a linear prediction model of the vocal tract, or by the envelope of the DFT spectrum of speech or the output of a set of filter-banks.

(b) Speech formants, including formant frequencies and bandwidth, and their trajectories in time. Formants are the resonance frequencies of vocal tract cavity; where the spectral envelope's peaks occur. Note

that the formant information is contained in the spectral envelope or in the linear prediction model of speech.

(c) The fundamental frequency of the opening and closing of glottal cords, i.e. the pitch.

(d) The temporal dynamics of speech parameters namely the time-variation of the spectral envelope, the formants, and the pitch.

(e) Intonation signals. Intonation signals are conveyed by the temporal dynamics of pitch across a segment of speech.

(f) Stress/emphasis patterns which are functions of pitch intonation, duration and energy.

The various commonly used speech models, namely linear prediction models are harmonic noise models are introduced next. The cepstrum features of speech are discussed in section 13.x on speech rcognition.

13.4 Linear Prediction Models of Speech

A widely used source-filter model of speech is the linear prediction (LP) model introduced in detail in Chapter x. LP models are used for speech coding, recognition and enhancement. A LP model is expressed as

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + e(m) \qquad (13.2)$$

where $x(m)$ is speech signal, a_k are the LP parameters and $e(m)$ is speech excitation. Note that the coefficients a_k model the correlation of each sample with the previous P samples whereas $e(m)$ models the part of speech that cannot be predicted from the past P samples.

In the frequency domain Equation 13.2 becomes

$$X(f) = \frac{E(f)}{1 - \sum_{k=1}^{p} a_k e^{-j2\pi fk}} = \frac{E(f)}{A(f)} = \frac{G \cdot U(f)}{A(f)} \qquad (13.3)$$

where $X(f)$ is the speech spectrum, $E(f)$ is the spectrum of excitation, $U(f)$ is the same as $E(f)$ but with normalised power, G is a gain factor and $G/A(f)$ is the spectrum of the LP model of the combination of vocal tract and nasal cavities and lips as well as the spectral slope due to glottal pulse. In a source-filter LP model of speech the spectral envelop of speech is modelled

by the frequency response of the LP model $G/A(f)$ whereas the finer harmonic and random noise-like structure of the speech spectrum is modelled by the excitation (source) signal $E(f)$.

The model parameters $\{a_k \ k=1, \ldots \ P\}$ can be factorised and described in terms of a set of complex conjugate and real roots, the so-called poles of the model $(\rho_k \ k=1 \ldots P\}$. The poles are related to the resonance or formants of speech. The model parameters can also be expressed in terms of the reflection coefficients of a lattice model of speech as described in section XX. Figure (13.4.c) shows the frequency response of a linear prediction model of a speech sound.

MatLab LPCSpeechDemo()

13.4.1 Line Spectral Frequencies

The line spectral frequencies (LSF) are an alternative representation of linear prediction parameters. LSFs are used in speech coding, and in the interpolation and extrapolations of LP model parameters, for their good interpolation and quantisation properties. LSFs are derived as the roots of the following two polynomials:

$$P(z) = A(z) + z^{-(P+1)} A(z^{-1})$$

$$= 1 - (a_1 - a_P)z^{-1} - (a_2 - a_{P-1})z^{-2} - \cdots - (a_P - a_1)z^{-P} + z^{-P+1}$$

$$(13.4)$$

$$Q(z) = A(z) - z^{-(P+1)} A(z^{-1})$$

$$= 1 - (a_1 + a_P)z^{-1} - (a_2 + a_{P-1})z^{-2} - \cdots - (a_P + a_1)z^{-P}1z^{-P+1}$$

$$(13.5)$$

where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_P z^{-P}$ is the inverse linear predictor filter. Clearly $A(z) = [P(z) + Q(z)]/2$. The polynomial equations (13.4, 13.5) can be written in factorised form as

$$P(z) = \prod_{i=1,3,5,\cdots} (1 - 2\cos\omega_i z^{-1} + z^{-2})$$

$$(13.6)$$

$$Q(z) = \prod_{i=2,4,6,\cdots} (1 - 2\cos\omega_i z^{-1} + z^{-2})$$

$$(13.7)$$

where ω_i are the LSF parameters. It can be shown that all the roots of the two polynomials have a magnitude of one and they are located on the unit circle and alternate each other. Hence in LSF representation the parameter vector $[a_1, a_2, \dots a_P]$ is converted to LSF vector $[\omega_1, \omega_2, \dots \omega_P]$.

Figure (13.8) shows a segment of voiced speech together with poles of its linear predictor model and the LSF parameters.
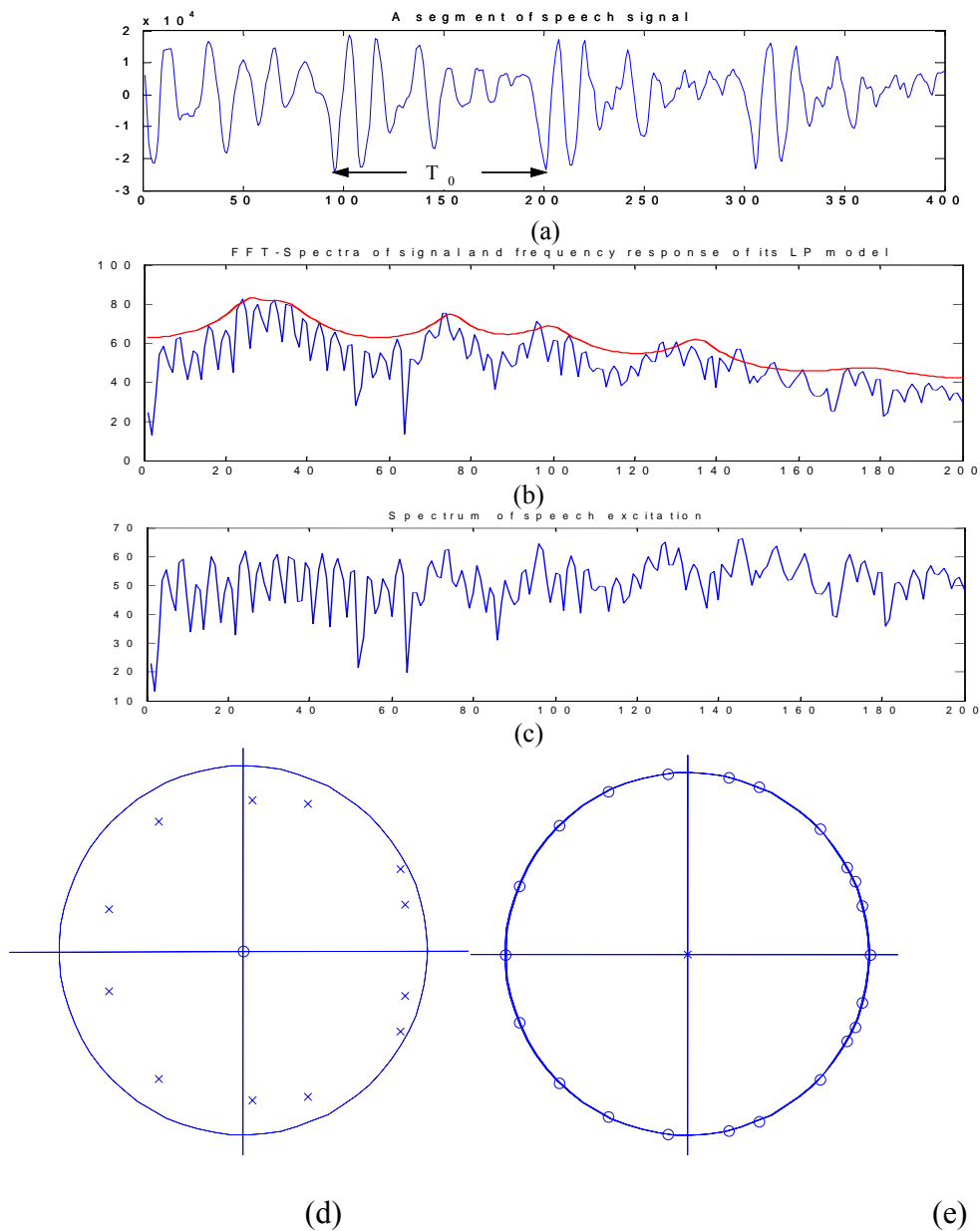

Figure 13.8 (a) A segment of speech signal, (b) its FFT and LP spectra, (c) the spectrum of its excitation, (d) the poles of its LP model, the roots of $P(z)$ and $Q(z)$ LSF polynomials.

13.5 Harmonic Pulse Noise Model of Speech

An alternative to a source-filter model of speech is the harmonic plus noise model. As the name implies speech signals can be represented by a combination of a harmonic model and noise model as

$$x(m) = \underbrace{\sum_{k=1}^{M} a_k(m)\cos\left(2\pi k F_0(m)m\right) + b_k(m)\sin\left(2\pi k F_0(m)m\right)}_{Harmonic\,Model\,(Fourier\,Series)} + \underbrace{v(m)}_{Noise\,Model} \quad (13.8)$$

where F_0 is the fundmental frequency, a_k, b_k are the amplitudes of the sine and cosine components of the kth harmonic, M is the number of harmonics and $v(m)$ is the noise-like random component of speech. Note that the harmonic part of the model is effectively a Fourier series respresentation of the periodic component of the signal.

In general for voiced signals the harmonic plus noise model of speech is composed of a mixture of both harmonics and noise. The proportion of harmonic and noise in voiced speech depends on a number of factors including the speaker characteristics (e.g. to what extent a speaker's voice is normal or breathy) ; the speech segment character (e.g. to what extent a speech segment is periodic) and on the frequency ; the higher frequencies of voiced speech have a higher poportion of noise-like components. Unvoiced speech are mostly composed of a spectrally-shaped noise-like component.

A segment of N samples of speech can be described in a vector-matrix notation as

$$\begin{bmatrix} x(m) \\ x(m-1) \\ x(m-2) \\ \vdots \\ x(m-N-1) \end{bmatrix} = \begin{bmatrix} \cos 2\pi F_0 m & \cdots & \cos 2\pi MF_0 m & \sin 2\pi F_0 m & \cdots & \sin 2\pi MF_0 m \\ \cos 2\pi F_0(m-1) & \cdots & \cos 2\pi MF_0(m-1) & \sin 2\pi F_0(m-1) & \cdots & \sin 2\pi MF_0(m-1) \\ \cos 2\pi F_0(m-2) & \cdots & \cos 2\pi MF_0(m-2) & \sin 2\pi F_0(m-2) & \cdots & \sin 2\pi MF_0(m-2) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos 2\pi F_0(m-N-1) & \cdots & \cos 2\pi MF_0(m-N-1) & \sin 2\pi F_0(m-N-1) & \cdots & \sin 2\pi MF_0(m-N-1) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_M \\ b_1 \\ \vdots \\ b_M \end{bmatrix} + \begin{bmatrix} v(m) \\ v(m-1) \\ v(m-2) \\ \vdots \\ v(m-N-1) \end{bmatrix}$$

$$(13.9)$$

In compact notation Equation (13.9) can be written as

$$x = Sc + v \qquad\qquad (13.10)$$

where x is the vector discrete-time speech samples, S is a matrix of sine and cosine functions, $c=[a\ b]$ is the vector of amplitudes of the harmonics and v is the noise component of the speech model. The harmonics amplitude vector c can be obtained from a least squared error minimisation process. Define an error vector as the difference between speech and its harmonic model as

$$e = x - Sc \qquad (13.11)$$

The squared error function is given by

$$ee^T = (x - Sc)(x - Sc)^T \qquad (13.12)$$

Minimistion of Equation(13.12) with respect to the amplitudes vector c yields

$$c = \left[S^T S \right]^{-1} S^T x \qquad (13.13)$$

Note that the harmonic amplitudes can also be obtained from the tracking of the peak amplitudes of the DFT of speech at the neigborhoods around the integer multiples of the fundamantal frequency F_0.

13.6 Fundamental Frequency (Pitch) Information

The periodic opening and closing of the vocal folds results in the harmonic structure in voiced speech signals. The inverse of the period is the fundamental frequency of speech. Pitch is the sensation of the fundamental frequency of the pulses of airflow from the glottal folds. The terms pitch and fundamental frequency of speech are used interchangeably.

The pitch of the voice is determined by four main factors. These include the length, tension, and mass of the vocal cords and the pressure of the forced expiration also called the sub-glottal pressure. The pitch variations carry most of the intonation signals associated with prosody (rhythms of speech), speaking manner, emotion, and accent. Figure (13.9) illustrates an example of the variations of the trajectory of pitch over time.

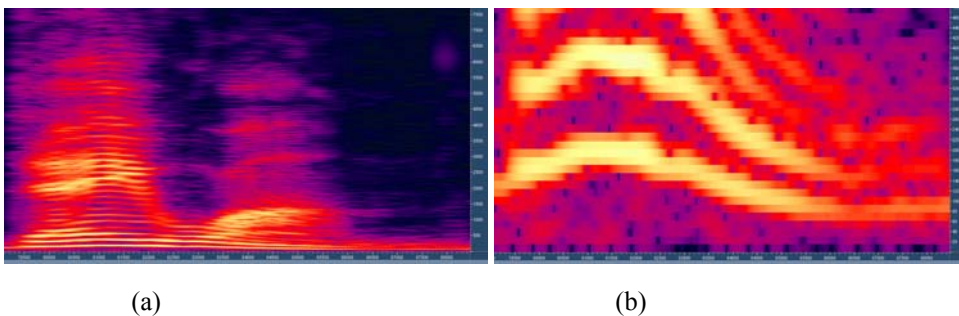   (a)   (b)

Figure 13.9 - (a) The spectrogram of the utterance *"day one"* showing the pitch and the harmonic structure of speech, (b) a zoomed spectrogram of the fundamental and the second harmonic of pitch.

The following information is contained in the pitch signal:

(a) Gender is conveyed in part by the vocal tract characteristics and in part by the pitch value. The average pitch for females is about 200 Hz whereas the average pitch for males is bout 110 Hz. Hence pitch is a main indicator of gender.

(a) Emotion signals in voice, such as excitement, stress, is also in part carried by pitch variations. Shouting as a means of stressing a point or in expression of anger has more to do with raising the pitch than loudness. Pitch variation is often correlated with loudness variation. Happiness, distress and extreme fear in voice are signalled by flutuations of pitch.

(b) Accent is in part conveyed by changes in the pitch and rhythm of speech. For example in some accents, such as the Northern Ireland accent, at the end of a sentence the pitch signal is raised instead of being lowered.

(c) Prosody is the rhythmical orintonation signals in speech carried mostly by the time-variations of pitch. The functions of prosody are many. Prosody can indicate syntax, demarcation and linking of phrases and sentences, turn-taking in conversational interactions, types of utterance such as questions and statements, and people's attitudes and feelings.

(d) Age and state of health. Pitch can also signal age, weight and state of health. For example children have a high pitched signal of 300-600 Hz.

13.6.1 Fundamental Frequency (Pitch) Estimation

Traditionally the fundamental frequency (whose sensation is known as pitch) is derived from the autocorrelation function as the inverse of the autocorrelation lag corresponding to the second largest peak of the autocorrelation function. Figure 13.10 shows a segment of voiced speech and its autocorrelation function. Note that the largest peak happens at the lag zero and corresponds to the signal energy. For a periodic voiced speech signal the second largest peak occurs at the lag T_0 corresponding to the period of speech.

The autocorrelation of a periodic signal is periodic with a period equal to that of the signal. Hence all the periodic peaks of the autocorrelation function can be usefully employed in the pitch estimation

process as in Griffin's methods [xx] where the pitch period is found by searching for the value of period T that maximises the following energy function

$$E(T) = T \sum_{k=0}^{N(T)} r(kT) \tag{13.14}$$

where $r(kT)$ is the autocorrelation at lag kT and $N(T)$ is the number of autocorrelation values in the summation. Note that the multiplication of the summation in Equation (13.14) by T compensates for the fact that as T
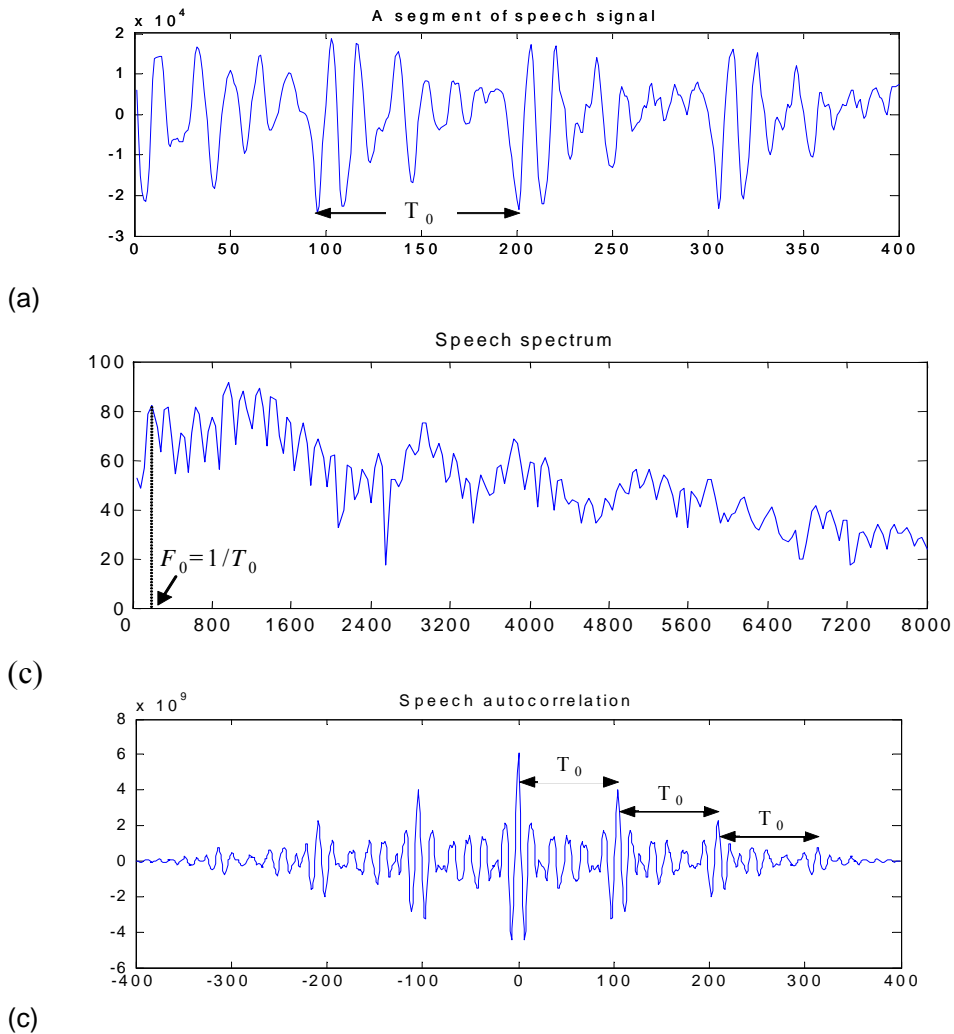

(a)

(c)

(c)

Figure 13.10 - (a) a voiced speech segment, (b) its frequency spectrum showing the harmonic structure (not sampling frequency was 16 kHz) and (c) its autocorrelation function peaks at integer multiples of the period of voiced speech.

increases the number of autocorrelation values in the summation, $N(T)$, decreases. The pitch is obtained as the maximum of Equation (13.14) as

$$T_0 = \arg \max_T E(T) \tag{13.15}$$

Figure 13.12 shows an example of the variation of $E(T)$ curve with a range of the values of periods. For each speech frame N pitch candidates are obtained as the N minimum values of $E(T)$ calculated on a grid of values of $T_{0min} < T_0 < T_{0max.}$. The Viterbi algorithm is subsequently used to obtain the best pitch trajectory estimate through the given N candidates. Figure 8 shows an example of speech and pitch ($F_0 = 1/T_0$) and harmonic tracks.

The pitch estimation method can be formulated in the frequency
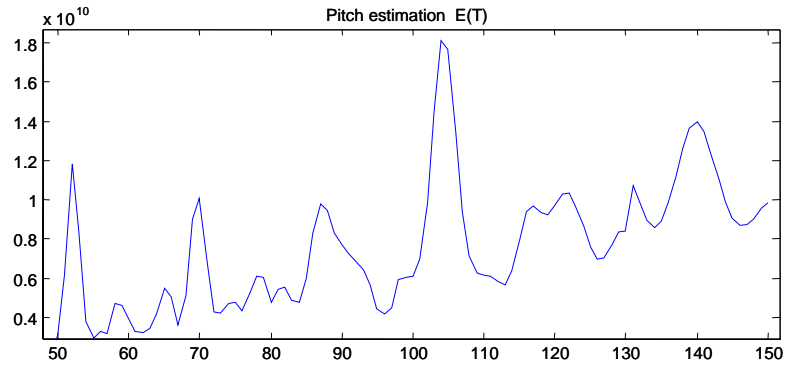

Figure 13.11- An illustration of the variation of $E(T)$ curve with the proposed values of pitch frequency T.
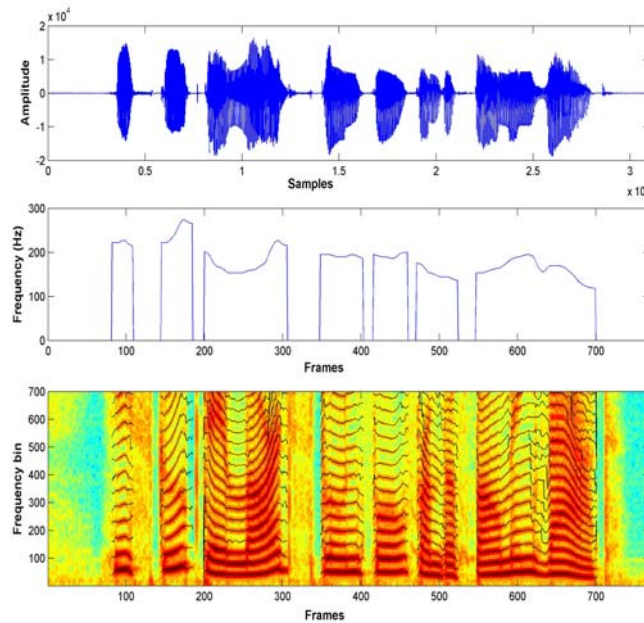

Figure 13.12 - An illustration of pitch tracks of a speech segment at sampling frequency of 8kHz.

domain to search for maximum signal to noise ratio at the harmonics.

A pitch estimation error criterion over the speech harmonics can be defined as

$$E(F_0) = F_0 \sum_{k=1}^{MaxF} \sum_{l=kF_0-M}^{kF_0+M} W(l) \log |X(l)| \qquad (13.16)$$

where $X(l)$ is the DFT of speech, F_0 is a proposed value of the fundamental frequency (pitch) variable and $2M+1$ is a band of values about each harmonic frequency. The use of the logarithmic compression in Equation (13.16) provides for a more balanced influence, on pitch estimation, of the high-energy low-frequency harmonics and the low-energy high-frequency harmonics. The weighting function $W(l)$ is a SNR-dependent Wiener-type weight given by

$$W(l) = \frac{SNR(l)}{1 + SNR(l)} \qquad (13.17)$$

where SNR(l) is the signal to noise ratio as the proposed l^{th} harmonic.

13.7 Speech Coding

The objective of speech coding is to achieve savings in the required memory storage space, transmission bandwidth and transmission power by reducing the number of bits per sample such that the decoded (decompressed) speech is perceptually indistinguishable from the original speech.

High quality speech coders can achieve a big reduction in the bit rate by factor of 13 or more (e.g. from 13 bits per sample to 1 bit per sample or less) with no perceptible loss in quality or intelligibility. Specifically speech coding methods achieve the following gains:

- A reduction in speech bit rate r_b, or equivalently the same reduction in the bandwidth ($BW=k.r_b$) and the memory storage requirement both of which are directly proportional to the bit rate.

- A reduction in the transmission power requirement because after compression there are less bits (hence less energy) per second to transmit.

- Immunity to noise, as error control coding methods can be used to reintroduced some of the saved bits per sample in order to protect speech parameters from channel noise and distortion.

Speech coding methods achieve a reduction in the bit rate by utilising the physiology of speech production and the psychoacoustics of audio perception, namely:

- Speech is a correlated signal; from an information theory point of view successive speech samples contain a high level of common or redundant information. The "redundancy" in natural speech provides pleasant sounds and robustness to background noise, speaker variations and accents. However, speech redundancy can be modelled and removed before transmission and then reintroduced into speech at the receiver.

- Speech is generated by a relatively slowly-varying articulatory system. Therefore, speech model parameters vary slowly and can be efficiently coded.

- Psychoacoustics of hearing. The threshold of hearing and the spectral and temporal noise masking thresholds can be used to allocate *just* sufficient number of bits to each subband in each speech frame to keep the coding noise masked below the thresholds of hearing.

13.7.1 Linear Prediction and Harmonic Noise Models in Speech Coding

Linear prediction model and harmonic noise model are the two main methods for modelling and coding of speech signals. Linear prediction model is particularly good at modelling the spectral envelop of speech whereas harmonic noise model is good at modelling the fine structure of speech. The two methods can be combined to take advantage of their relative strengths. In general, the main structures used for speech coding are:
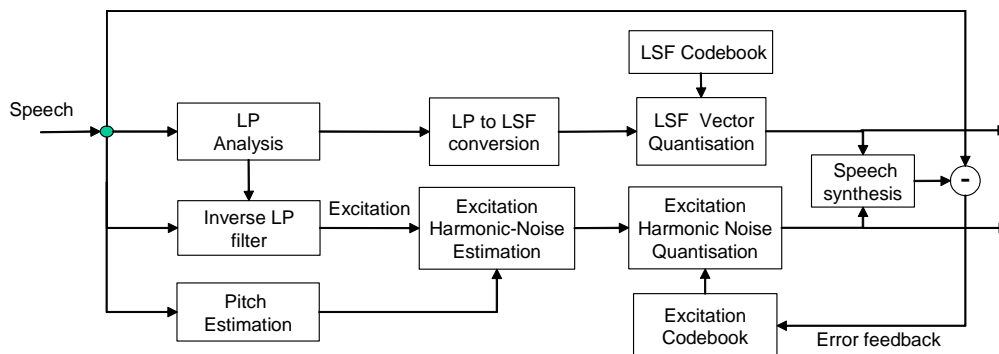

Figure 13.13 - An illustration of the outline of a speech coding system.

(a) The spectral envelop of speech modelled by a LP model and represented by a set of LSF coefficients.

(b) The speech excitation modelled by a harmonic and noise model.

Figure 13.13 shows the outline of a model-based speech coder. The speech signal is analysed and its LP model, excitation signals and pitch are extracted. The LP model represents the spectral envelop of speech. It is converted to a set of LSF coefficients which have good quantisation properties. The LSF coefficients can be scalar quantised or more efficiently they can be vector quantised using previously trained LSF vector codebooks.

The most challenging part of speech coding is the coding of the harmonic and noise contents of excitation. To start with, unlike the extraction of LP model parameters, the extraction of the fundamental frequency is not a straight forward matter; whereas there is a closed-form solution for extraction of LP model parameters (see Chapter xx) the calculation of pitch and harmonics usually requires a search process. The main issues on modelling the harmonic and noise parts of speech excitation are the following:

(1) Voiced/Unvoiced classification..

(2) Pitch estimation.

(3) Estimation of harmonic amplitudes.

(4) Estimation of noise variance

(5) Quantisation of harmonic and noise parameters.

In its simplest form one can model speech excitation as a mixture of harmonics plus noise. More sophisticated methods employ different mixtures of harmonic and noise for each speech subband centred on a harmonic.

13.7.2 Modelling Excitation for Voiced Speech

The excitation signal for voiced speech is composed of both harmonic and noise. The proportion of harmonic and noise in the mixture around each harmonic frequency depends on the speaker, phoneme and frequency. Generally, the proportion of noise to harmonic content of speech increases with the frequency. Equation (xx) can also be used to describe the excitation

for voiced speech in terms of the Fourier series of a periodic component plus a noise component as

$$e(m) = \left[a_k(m)\cos\left(2\pi k F_0(m)m\right) + b_k(m)\sin\left(2\pi k F_0(m)m\right) \right] + v(m) \quad (13.18)$$

Alternatively, we may express the excitation signal in the complex frequency domain as

$$E(f) = \sum_{k=1}^{N} A_k M(f - kF_0) + V(f) \quad (13.19)$$

where $E(f)$ is the excitation, F_0 is the fundamental frequency, A_k is the complex amplitude of the kth harmonic, $M(f)$ is a harmonic shape function that may be a delta function (for a sine wave) or a Gaussian function with unit area, and $V(f)$ is the noise component of the excitation. Note that since speech is a real-valued signal, for each complex value of $E(f)$ we also have its symmetric complex conjugate value. An estimate of A_k can be obtained from tracking the peaks of the DFT of speech.

We can define a harmonicity parameter for the kth harmonic as the proportion of the harmonic and noise components in a subband of bandwidth F_0 centred at the kth harmonic (i.e. $kF_0 - F_0/2 \leq f \leq kF_0 + F_0/2$). The harmonicity may be defined as

$$H_k = 1 - \frac{\displaystyle\int_{-F_0/2}^{F_0/2} \left[\left| E\left(kF_0\right) \right| M\left(f - kF_0\right) - \left| E\left(f - kF_0\right) \right| \right]^2 df}{\displaystyle\int_{-F_0/2}^{F_0/2} \left| E\left(f - kF_0\right) \right|^2 df} \quad (13.20)$$

where $E(kF_0)=|A_k|$. Note that H_k is the proportion of harmonic and the proportion of noise in the harmonic noise model is given by

$$V_k = 1 - H_k = \frac{\displaystyle\int_{-F_0/2}^{F_0/2} \left[\left| E\left(kF_0\right) \right| M\left(f - kF_0\right) - \left| E\left(f - kF_0\right) \right| \right]^2 df}{\displaystyle\int_{-F_0/2}^{F_0/2} \left| E\left(f - kF_0\right) \right|^2 df} \quad (13.21)$$

The harmonicity values can be used to reconstruct the proportion of harmonic and noise in each speech subband.

13.7.3 Modelling Excitation for unvoiced speech

The unvoiced part of the excitation is a noise-like signal with a white spectral envelop. Indeed the unvoiced part of the excitation can be a Gaussian noise; what are important in the coding and reconstruction of unvoiced speech segments are the shapes of the spectral envelops and the gain factors i.e. the power of the noise. Hence in the coding of unvoiced speech segments we only need to transmit LSF parameters of the LP model and the gain and then any Gaussian noise of unit variance would produce a perceptually high quality signal.

13.7.4 Code-Excited Linear Prediction Coding

For applications such as mobile phones, speech coders utilise a form of code-excited linear prediction (CELP) model of speech. Figure (13.14) shows an outline of a CELP coder. As shown CELP models speech as the output of filter excited by a combination of a periodic component and a non-periodic component.

Outline of Principles of Operation of CELP

The principles of the operation of CELP coders is as follows:

(1) *Sampling, PCM Quantisation and Segmentation.* Before CELP coding, the input signal to the handset's microphone is filtered and sampled at a rate of 8000 samples per second. Each sample is then quantised with 13 bit per sample. The sampled speech is segmented into segments of 20 ms (160 samples) long.

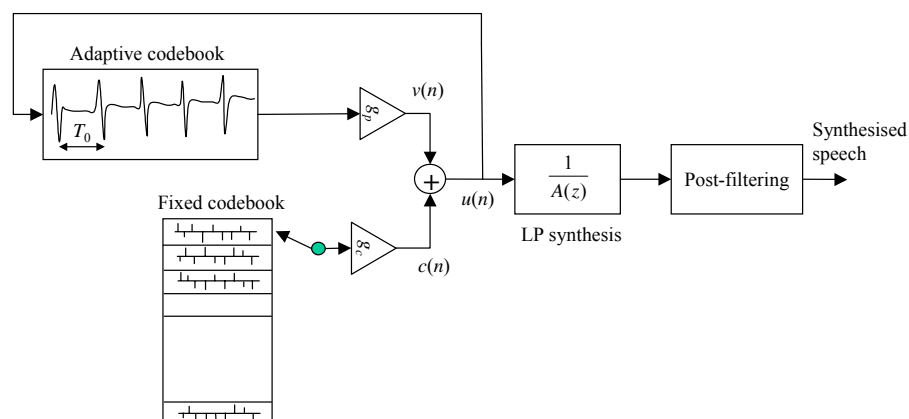(2) *Linear Prediction Analysis.* Each speech segment is windowed and

Figure 13.14 The outline of a simple code excited linear prediction (CELP) decoder.

modelled by a 10th order linear prediction model. The z-transfer function of the linear prediction filter is given as

$$H(z) = \frac{1}{1 + \displaystyle\sum_{k=1}^{P} a_k z^{-k}} \qquad (13.22)$$

The linear predictor coefficients are calculated using the Levison-Durbin method. The predictor coefficients are transformed into line spectral frequencies before quatisation and transmission.

(3) *Calculation of the Periodic Part of Speech.* The periodic part of speech is syntesised by exciting the linear predictor filter with a periodic input obtained from an adaptive codebook. A pitch filter is used to shape the periodic component of excitation. The pitch synthesis filter model has the following transfer function

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T}} \qquad (13.23)$$

where T is the pitch period and g is pitch filtr coefficient. For each speech segment the pitch period T is calculated from the autocorreation function of speech.

(4) *Calculation of the Non-Periodic Part of Speech.* The non-periodic part of speech is obtained by exciting the linear prediction model with the 'best' noise-like excitation selected from a number of available 'noise codes' using an analysis-by-synthesis search method. The anslysis by synthesis approach test all possible input forms available and selects the one that minimises the perceptually weighted mean squared difference between the original speech and the synthesised speech. This analysis by synthesis part of the coding can also compensate for the deficiences in the earlier part ofs of the coding process.

(5) The reconstructed speech is passed through an adaptive post-filter.

Details of Operation of CELP

Figure (13.15) shows a more detailed block diagram of a code excited linear prediction (CELP) coder used in mobile phones, The analogue speech input
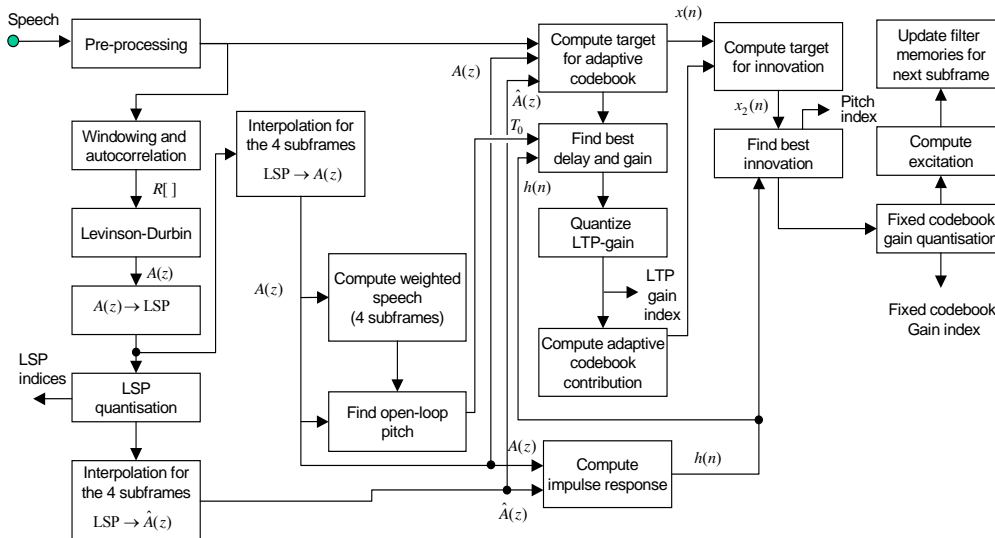
Figure 13.15 Block diagram of a code excited linear prediction (CELP) coder.

is sampled at 8 kHz with an 8-bits A-law device. The 8-bits A-law format is converted to a 13-bits linear format using the method and tables defined by ITU-T Rec. G.713. The incoming samples are divided into frames of 160 samples (i.e. 20 ms at 8 kHz) during which speech parameters are assumed time-invariant. The frame length of 20 ms plus the signal processing time on the handset and the delay at the network determine the "trans-coding delay" of the communication system. The encoder compresses an input frame of 160 13-bits PCM samples (a total of 2080 bits/frame) to a frame of 260 bits. This is a compression ratio of 8:1. The speech coder of Figure (13.15) consists of the following sections.

13.7.4.1 Pre-processing High-pass Filter

This is a second order high-pass filter for removing unwanted low frequency components below 80 Hz. The filter transfer function shown in Figure (13.10) is given by

$$H(z) = \frac{0.92727435 - 1.8544941z^{-1} + 0.92727435z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}} \qquad (13.24)$$

13.7.4.2 Linear Prediction Model Analysis and Quantisation

The steps taken in the calculation and quantisation of the short-term linear prediction coefficients, shown in Figure (13.9) are as follows.
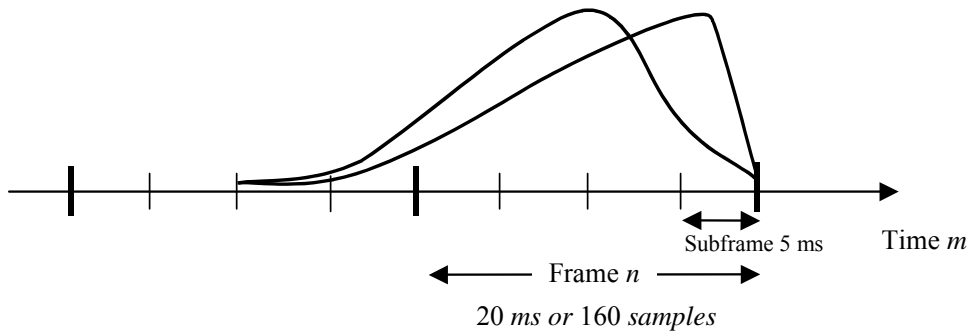
Figure 13.17 Linear prediction analysis windows.

Windowing and autocorrelation: Linear prediction analysis is performed twice per frame with two asymmetric windows. Each window covers one and a half frame of speech to provide some overlap of the signal between the current and the previous frames. One window has its weight concentrated on the second speech sub-frame and is defined as

$$w_I(m) = \begin{cases} 0.54 - 0.46\cos\left(\pi m / \left(L_1^{(I)} - 1\right)\right) & m = 0, \ldots, L_1^{(I)} - 1 \\ 0.54 + 0.46\cos\left(\pi(m - L_1^{(I)}) / \left(L_2^{(I)} - 1\right)\right) & m = L_1^{(I)}, \ldots, L_1^{(I)} + L_2^{(I)} - 1 \end{cases}$$

$$(13.25)$$

Figure 13.16 The frequency response of the pre-processing filter.

where $L_1^{(I)} = 160$ and $L_2^{(I)} = 80$. The second window has its weight concentrated in the fourth subframe and is given by

$$
w_{II}(m) = \begin{cases} 0.54 - 0.46\cos\left(2\pi m \big/ \left(L_1^{(II)} - 1\right)\right) & m = 0, \ldots, L_1^{(II)} - 1 \\ \cos\left(2\pi (m - L_1^{(II)}) \big/ \left(4L_2^{(II)} - 1\right)\right) & m = L_1^{(II)}, \ldots, L_1^{(II)} + L_2^{(II)} - 1 \end{cases}
$$
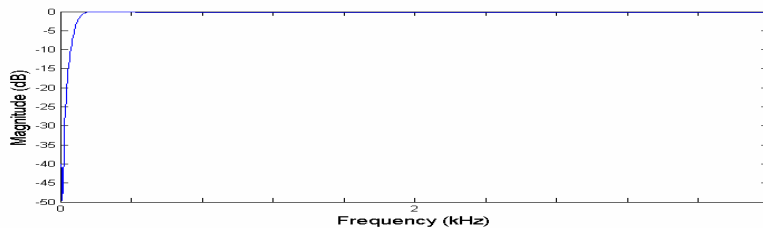
(13.26)

where $L_1^{(II)} = 232$ and $L_2^{(II)} = 8$. The windows span 240 samples including 160 samples from the current frame and 80 samples from the previous frame. From each windowed speech segment $P+1$ autocorrelation coefficients are calculated and used to obtain P LP coefficients.

The windows are given in the following Matlab code.

```
function [wI]=GSM_WindowI()
L1=160; L2=80;
m=1:L1;  wI(m)=0.54-0.46*cos(pi*(m-1)/(L1-1));
m=L1+1:L1+L2;  wI(m)=0.54+0.46*cos(pi*(m-L1-1)/(L2-1));
function [wII]=GSM_WindowII()
L1=232; L2=8;
m=1:L1; wII(m)=0.54-0.46*cos(2*pi*(m-1)/(2*L1-1));
m=L1+1:L1+L2; wII(m)=cos(2*pi*(m-L1-1)/(4*L2-1));
```

LP coefficients calculation: The Levinson-Durbin algorithm, described in Section (xxx), is used to obtain P LP coefficients $[a_1, a_2, \ldots a_P]$ from $P+1$ autocorrelation coefficients $[r(0), r(1), \ldots r(P)]$. Alternatively the Schur algorithm can be used.

LP to LSF conversion and quantisation: The LP coefficients are converted to LSF coefficients, described in Section (xxx), LSFs have good quantisation and interpolation properties. The LSF quantisation is achieved as follows. First the two sets of LP coefficient vectors from the two windows per speech frame are quantified using the LSF representation in the frequency domain as

$$
f_i = \frac{F_s}{2\pi} \arccos(\omega_i)
$$

(13.27)

where f_i is the line spectral frequency (LSF) in Hz, F_s is the sampling frequency (8 kHz for telephones) and ω_i are the line spectral angular frequencies. The LSF vector is $\boldsymbol{f}=[f_1, f_2, \ldots, f_P]$. The prediction and quantisation of the LSF vectors are performed as follows. Let $\boldsymbol{z}^{(1)}(m)$ and $\boldsymbol{z}^{(2)}(m)$ denote the mean-removed LSF vectors at frame m. The prediction LSF residual vector is defined by

$$\boldsymbol{r}^{(1)}(m) = \boldsymbol{z}^{(1)}(m) - \boldsymbol{p}(m) \tag{13.28}$$

$$\boldsymbol{r}^{(2)}(m) = \boldsymbol{z}^{(2)}(m) - \boldsymbol{p}(m) \tag{13.29}$$

where $\boldsymbol{p}(m)$ is the predicted LSF vector at frame m given by a first order moving-average predictor as

$$\boldsymbol{p}(m) = \alpha\,\boldsymbol{r}^{(2)}(m-1) \tag{13.30}$$

where α=0.65. The two LSF vectors $\boldsymbol{r}^{(1)}(m)$ and $\boldsymbol{r}^{(2)}(m)$ are quantised using split matrix quantisation (SMQ). For a 10th-order linear prediction model the LSF residual matrix $[\boldsymbol{r}_1^{(1)}(m)\ \boldsymbol{r}_2^{(2)}(m)]$ is split into five 2×2 matrices. For example, the first submatrix will be $[r_1^{(1)}(m)\ r_1^{(2)}(m)\ r_2^{(1)}(m)\ r_2^{(2)}(m)]$. The five submatrices are then vector quantised using a bit allocation pattern of [7, 8, 9, 8, 6]. Note the total number of bits per frame alocated to LSF coefficients is 38 bits per 20 ms frame. In general an input LSF vector $\boldsymbol{f}$ is quantised using the following VQ error relation:

$$E = \sum_{i=1}^{10}\left[w_i\left(f_i - \hat{f}_i\right)\right]^2 \tag{13.31}$$

where w_i are weighting coefficient.

Interpolation of LSF parameters

The LSF parameters are interpolated in the angular frequency domain ω. The LSF coefficients calculated for the windows centred on the second and the fourth subframes are used for those subframes respectively. The LSF coefficients for the first and the third subframes are linearly interpolated from the adjacent subframes as

$$\boldsymbol{\omega}^{(1)}(m) = 0.5\boldsymbol{\omega}^{(4)}(m-1) + 0.5\boldsymbol{\omega}^{(2)}(m) \tag{13.32}$$

$$\boldsymbol{\omega}^{(3)}(m) = 0.5\boldsymbol{\omega}^{(2)}(m) + 0.5\boldsymbol{\omega}^{(4)}(m) \tag{13.33}$$

where $\omega^{(i)}(m)$ is the LSF coefficients vector of the ith subframe of the mth speech frame.

Open-loop pitch analysis

The period T of a periodic signal can be obtained from an analysis of the maxima of its autocorrelation function. Open-loop pitch estimation is performed every 10 ms, i.e. twice per frame, from the autocorrelation function

$$r(k) = \sum_{m=0}^{79} x(m)x(m-k) \tag{13.34}$$

At each stage the three maxima of the autocorrelation function are obtained in the following ranges: $k=18\text{-}35$, $k=36\text{-}71$ and $k=72{:}143$. The retained maxima are normalised by $\sqrt{\sum_m x^2(m-t_i)}$, where t_i are the maxima. The normalised maxima and corresponding delays are denoted as (M_i, t_i) $i=1,2,3$. The following method is then used to select the pitch from the three candidates

$T_{op} = t_1$
$M(T_{op}) = M_1$
if $M_2 > 0.85M(T_{op})$
   $M(T_{op}) = M_2$
   $T_{op} = t_2$
end
if $M_3 > 0.85M(T_{op})$
   $M(T_{op}) = M_3$
   $T_{op} = t_3$
end

The above procedure of computing the maxima of the correlation function in three ranges, and its bias in favouring the lower range, is designed to avoid choosing pitch multiples.

Adaptive codebook search: The adaptive codebook search is performed every 5 ms for each speech subframe to obtain the pitch parameters i.e. the pitch delay and the pitch gain. The pitch values are optimised in a closed-loop pitch analysis performed around the open-loop pitch estimates by minimising the difference between the input speech and the synthesised

Track	Pulse	Positions
1	i_0, i_5	0, 5, 10, 15, 20, 25, 30, 35
2	i_1, i_6	1, 6, 11, 16, 21, 26, 31, 36
3	i_2, i_7	2, 7, 12, 17, 22, 27, 32, 37
4	i_3, i_8	3, 8, 13, 18, 23, 28, 33, 38
5	i_4, i_9	4, 9, 14, 19, 24, 29, 34, 39

Table 13.3- Potential positions of the individual pulses in the algebraic code.

speech. For the first and the third subframes the range $T_{op}\pm3$ bounded by 18…143 is searched. A fractional pitch delay is used with a resolution of 1/6 in the range [$17\frac{3}{6}$, $94\frac{3}{6}$] and integers only in the range [65, 143]. For the second and fourth subframes, a pitch resolution of 1/6 is always used in the range [$T_1-5\frac{3}{6}$, $T_1+4\frac{3}{6}$], where T_1 is the nearest integer to the fractional pitch lag of the previous (1st or 3rd) subframe. The pitch delay is coded with 9 bits in the first and third subframes and the relative delay of the other frames is coded with 6 bits.

Algebraic codebook structure and search

The algebraic code structure is based on interleaved single-pulse permutation (ISPP) method. Each codebook vector of size 40 samples contains 10 non-zero pulses with amplitudes ±1. Each subframe of 40 samples is subdivided into five tracks, where each track contains two pulses as shown in table 2. Each two pulse positions in an eight-positions track is coded with 3 bits (a total of 6 bits/track) and the sign of the first pulse in each track is coded with one bit (a total of 5 bits per subframe). The sign of the second pulse is the opposite of the first pulse if its position is smaller than the first pulse but it has the same sign as the first pulse otherwise. This gives a total of 35 bits for each 40-samples subframe which are Grey coded for robustness. The algebraic codebook is searched for the best vector by minimising the difference between the input speech and the synthesised speech.

The use of 'excitation noise' to compensate for modelling errors. Note that in speech coders with a closed loop error minimisation system the choice of the noise compensates for the inadequacies and errors in the modelling of LP parameters, pitch and harmonic values. Hence, a main reason that much processing power and bit rate is used to obtain and

transmit the best noise sequence is because the 'best noise' sequence also compensates for other errors by trying to match the coded speech with the actual speech in the error minimisation feedback loop as explained in the next section.

Calculation of Periodic amd Non-Periodic Excitation Gain Factors . The gain factor for the synthesised periodic part of speech $y(m)$ is calculated as the normalised correlation of the synthesised periodic part of speech and the actual speech $x(m)$ as

$$g_p = \frac{\sum_{n=0}^{N-1} x(m)y(m)}{\sum_{n=0}^{N-1} y(m)y(m)} \tag{xx}$$

Similarly the gain for the nonperiodic synthesised part of speech can be calculated as

$$g_c = \frac{\sum_{n=0}^{N-1} x_2(m)z(m)}{\sum_{n=0}^{N-1} z(m)z(m)} \tag{xx}$$

where $x_2(m)$ is the differene between the actual speech and its synthesised perodic part and $z(m)$ is the the synthesised nonperiodic part of speech. That is $z(m)$ is the output of the linear prediction filer in response to the selected noise from the codebook.

Parameter	1st & 3rd subframes	2nd & 4th subframes	Total bits per frame
2 LSF sets	19		38
Pitch delay	9	6	30
Pitch gain	4	4	16
Algebraic codebook	35	35	140
Codebook gain	5	5	20
Total			**240**

Table 13.3- Bit Allocation pattern in a 12.2 kbps Celp coder

13.5.4 Decoding and Re-synthesis of Coded Speech

At the receiver the decoding and re-synthesis of speech is performed as follows.

Decoding of LP filter coefficients: The received codebook indices for LSP quantisation are used to reconstruct the LSP coefficients for the second and fourth subframes. The LSP coefficients for the first and the third subframes are derived from interpolation of the adjacent LSP vectors as described in Equations (13.26) and (13.27). The LSPs are then converted to LP vectors. Then the following steps are repeated for each subframe:

(6) Decoding of the adaptive codebook vector: The received pitch index is used to find the integer and fractional parts of the pitch lag. The adaptive codebook vector $v(m)$ is obtained by interpolating the past excitation $u(\text{m})$ at the pitch delay.

(7) Decoding of the adaptive codebook gain. The received index is used to find the quantised adaptive codebook gain g_p from the quantisation table.

(8) The received algebraic codebook index is used to extract the positions and amplitude signs of the excitation pulses and to find the algebraic code vector $c(m)$.

(9) Decoding of the fixed codebook gain. The received index is used to compute the quantised fixed codebook gain g_c.

Reconstructing Speech

The input excitation to linear prediction filter is computed as Figure (13.8)

$$u(n) = g_p v(n) + g_c c(n) \qquad (13.35)$$

The input excitation $u(m)$ is filtered by the LP filter $1/A(z)$ and then processed by an adaptive postfiltering filter as described next .

Adaptive Post-filtering

The adaptive post-filter is the cascade of two filters: a formant post-filter and a tilt compensation filter. The postfilter is updated every subframe of 5 ms. The formant postfilter is given by:

$$H_f(z) = \frac{\hat{A}(z/\gamma_n)}{\hat{A}(z/\gamma_d)} \tag{13.36}$$

where $\hat{A}(z)$ is the received quantized and interpolated LP inverse filter and the factors γ_n and γ_d control the amount of the formant post-filtering. Finally the filter $H_t(z)$ compensates for the tilt in the formant post-filter and is given by

$$H_t(z) = (1 - \mu z^{-1}) \tag{13.37}$$

where $\mu = \gamma_t k_1$ is a tilt factor, with k_1 being the first reflection coefficient calculated on the truncated (L_h=22) impulse response, $h_f(m)$ of the filter Equation (13.36). The reflection coefficient k_1 is given by

$$k_1 = \frac{r_h(1)}{r_h(0)}; \qquad\qquad r_h(i) = \sum_{j=0}^{L_h-i-1} h_f(j) h_f(j+i) \tag{13.38}$$

The post-filtering is performed by first passing the synthesised signal through $\hat{A}(z/\gamma_n)$ then through $1/\hat{A}(z/\gamma_d)$ and finally through the tilt filter $H_t(z)$.
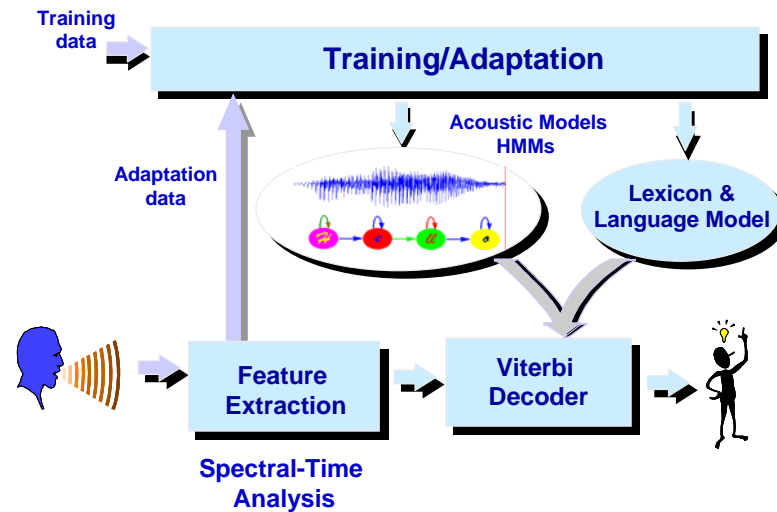
Figure 13.18 -The outline of a speech recognition system.

13.8 Speech Recognition

Speech recognition systems have a wide range of applications from the relatively simple isolated-word recognition systems for name-dialling, automated costumer service and voice-control of cars and machines to continuous speech recognition as in auto-dictation or broadcast-news transcription. Figure (13.18) shows the outline of a typical speech recognition system. It consists of three main sections:

(a) A front-end section for extraction of a set of spectral-temporal speech features from the time-domain speech samples. The speech features are derived from a bank of filers inspired by a knowledge of how the cochlear of the inner ear performs spectral analysis of audio signals. The most commonly used features are cepstrum features described later in this section.

(b) A middle section that consists of a network of speech models incorporating; pre-trained statistical models of the distributions of speech features, a language model and speaker adaptation. In its simplest form, e.g. for a name-dialling system, a speech model consists of a simple spectral-temporal template for each word. In more high performance systems a lattice network incorporates hidden Markov models of speech with language models.

(c) A speech decoder, usually based on a Viterbi decoding method, takes as input a stream of speech feature vectors and, using a

network of acoustic word models, outputs the most likely word sequence.

The speech recognition problem can be stated as follows: *given a stream speech features **X** extracted from the acoustic realisation of a spoken sentence, and a network of pre-trained speech models **Λ**, decode the most likely spoken word sequence **W**=[w₁,w₂, ..., wₙ].* Note that speech model network Λ contains models of acoustic features representation of words and can also include a model of language grammar. Formally, the problem of recognition of words from a sequence of acoustic speech features can be expressed in terms of maximisation of a probability function as

$$[\hat{w}_1, \hat{w}_2, ... \hat{w}_N] = \max_{[w_1, w_2, w_3, ... w_N]} f\left([w_1, w_2, ... w_N] \big| X, \Lambda\right) \qquad (13.39)$$

where $f(\cdot)$ is the conditional probability of a sequence of words **W** given a sequence of speech features **X**. Note that in practice speech is processed sequentially and the occurrence of a word, a phoneme or a speech frame is conditioned on the previous words, phonemes and speech frames.

The fundamental problems of speech recognition. Like any pattern recognition problem, the fundamental problem in speech recognition is the speech pattern variability. Speech recognition errors are caused by the overlap of the distributions of the acoustic realisations of different speech units. In general the sources of speech variability are as follows:

(a) **Duration variability.** No two spoken realisations of a word, even by the same person, have the same duration. Furthermore, often the variation in the duration of a word is non-uniform in that in different realisations of a word parts of a word may be more elongated/shrunk compared to other parts of the same word.

(b) **Spectral variability.** No two spoken realisations of a word, even by the same person, have the same spectral-temporal trajectory.

(c) **Speaker variability.** Speech is affected by the anatomical characteristics, gender, health, and emotional state of the speaker.

(d) **Accent.** Speaker accent can have a major effect on speech characteristics and on speech recognition performance.

(e) **Contextual variability.** The characteristic of a speech unit is affected by the acoustic and phonetic context of the units preceding or succeeding it.

(f) **Co-articulation.** This is similar to contextual variability but is also affected by the speech rate, accent and psychological factors.

(g) **Noise.** Speech recognition is affected by noise, echo, channel distortion and adverse environment.

Speech recognition methods aim to model, and where possible reduce, the effects of the sources of speech variability. The most challenging sources of variations in speech are speaker characteristics including accent, co-articulation and background noise.

Isolated-word and continuous-speech recognition. In terms of the 'fluency' or continuity of the input speech that a speech recognition system can handle, there are broadly two types of speech recognition systems:

(a) *Isolated-word recognition systems*, with short pauses between spoken words, are primarily used in small vocabulary command-control applications such as name-dialling, Internet navigation, and voice-control of computer menus or accessories in a car. Isolate-word recognition systems may use models trained on whole word examples or constructed from concatenation of sub-word models.

(b) *Continuous speech recognition* is the recognition and transcription of naturally spoken speech. Continuous speech recognition systems are usually based on word model networks constructed from compilation of phoneme models using a phonetic transcription dictionary. Usually the word model networks also include an *N*-gram language model.

Speaker-dependent and speaker-independent speech recognition.

In any communication system the decoding errors at the receiver increases with the increasing overlap of the probability distributions of the signals that carry the communication symbols. In a speaker-independent speech database, the effect of variations of different speakers' characteristics results in higher variances and hence broader probability distributions of speech features. Broader distributions of different acoustic speech sounds result in higher overlaps between those distributions and hence a higher speech recognition error rate.

Unlike speaker-independent systems, speaker-dependent systems do not have to deal with the extra variance due to the variations of speakers' characteristics and the consequent additional overlap of the probability distributions of different speech sounds. Hence, speaker-dependent systems can be more accurate than speaker-independent systems. However, the performance of speaker-independent systems can be improved using speaker adaptation methods, such as maximum likelihood linear regression (MLLR) method, to modify the parameters of a speaker-independent system towards those of a speaker-dependent system.

13.8.1 Speech Units

Every communication system has a set of elementary symbols (or alphabet) from which larger units such as words and sentences are constructed. For example in digital communication the basic alphabet are "1" and "0", and in written English the basic units are *A* to *Z*. The elementary linguistic unit of spoken speech is called a *phoneme* and its acoustic realisation is called a *phone*. There are between 60 to 80 phonemes in spoken English; the exact number of phonemes depends on the dialect. For the purpose of automatic speech processing the number of phonemes is clustered and reduced to between 40 to 60 phonemes depending on the dialect. The phonemes in American English and British English are listed in Tables 1 and 2 respectively. The British English phonemes are grouped into ten categories. The US English can be similarly classified.

Note that phonetic units are not produced in isolation and that their articulation and temporal-spectral "shape" is affected by the context of the preceding and succeeding phones as well as the linguistic, expressional and tonal context in which they are produced. For speech recognition context-dependent triphone units are used. A triphone is a phone in the context of a preceding and a following phones for example the triphones for the word *imagination* are *iy+m iy-m+ae m-ae+g ae-g+iy g-iy+n iy-n+ay where –* denotes preceding and + denotes succeeding.

Assuming that there are about 40 phonemes, theoretically there will be about 40×40=1600 context dependent variations for each phone, and hence a total of 40×1600=64000 triphones. However, due to linguistic constraints some triphones do not occur in practice.

Phoneme	Example	Phoneme	Example	Phoneme	Example	Phoneme	Example
IY	b**ea**t	IX	ros**es**	NX	si**ng**	V	**v**at
IH	b**i**t	ER	b**ir**d	P	**p**et	DH	**th**at
EY	b**ai**t	AXR	butt**er**	T	**t**en	Z	**z**oo
EH	b**e**t	AW	d**ow**n	K	**k**it	ZH	a**z**ure
AE	b**a**t	AY	b**uy**	B	**b**et	CH	**ch**urch
AA	b**o**b	OY	b**oy**	D	**d**ebt	JH	**j**udge
AH	b**u**t	Y	**y**ou	H	**g**et	WH	**wh**ich
AO	b**ou**ght	W	**w**it	HH	**h**at	EL	batt**le**
OW	b**oa**t	R	**r**ent	F	**f**at	EM	bott**om**
UH	b**oo**k	L	**l**et	TH	**th**ing	EN	butt**on**
UW	b**oo**t	M	**m**et	S	**s**at	DX	ba**tt**er
AX	**a**bout	N	**n**et	SH	**sh**ut	Q	(glottal stop)

Table 13.1 – A list of American English phonemes.

Vowels	Semivowel	Nasal	Vfricative	Whisper
aa	l	m	v	hh
ae	r	n	dh	sil
ah	w	ng	z	
ao	y	**Affricative**	zh	
ax	**Diphthong**	ch	**UnVstop**	
eh	aw	jh	p	
er	ay	**UnVfricative**	t	
ih	ea	f	k	
iy	ey	s	**Vstop**	
oh	ia	sh	b	
ow	oy	th	d	
uh	ua		g	
uw				

Table 13.2 – A categorised list of British English phonemes.

Syllables are also subword units, but they are larger than phonemes. A word may be composed of one or more syllables and a syllable may be composed of one or more phonemes. For example, the word '*imagination*' can be deconstructed into the following sub-word units:

Word	*imagination*
Phonetic transcription	*iy m ae g iy n ay sh e n*
Triphone transcription	*iy+m iy-m+ae m-ae+g ae-g+iy g-iy+n iy-n+ay*
	n-ay+sh ay-sh+ay sh-e+sh e-n+sh e-n
Syllable transcription	*iyma giy nay shen*

13.8.2 Entropy of speech

In information theory, entropy is defined as a measure of the randomness or information content of a process. Entropy quantifies the information content or the capacity of an information source, as described in Section (). The information in speech - due to such random variables, as words, speaker, intonation and accent - can be quantified in terms of entropy.

Entropy of an information source gives the theoretical lower bound for the number of binary bits required to encode the source. The entropy of a set of communication symbols is obtained as the probabilistic average of $\log_2$ of the probabilities of the symbols. The entropy of a random variable X with M states or symbols $X=[x_1, ..., x_M]$ and the state or symbol probabilities $[p_1, ..., p_M]$, where $P_X(x_i)= p_i$, is given by

$$H(X)=-\sum_{i=1}^{M}P_X(x_i)\log P_X(x_i)$$ (13.40)

We can consider the entropy of speech at several levels; such as the entropy of words contained in a sequence of speech or the entropy of intonation or the entropy of the speech signal features. The calculation of the entropy of speech is complicated as speech signals simultaneously carry various forms of information such as phonemes, topic, intonation signals, accent, speaker voice, speaker stylistics.

Example 13.1 Entropy of Phonemic Symbols - Speech is made up of about 40 basic acoustic symbols, known as phonemes, these are used to construct words, sentences etc. Assuming that all phonetic units are equiprobable, and that the average speaking rate is 10 phonemes/second, calculate the minimum number of bits per second required to encode speech at the average speaking rate. For speech $M=40$ assume $P(x_i)=1/40$, the entropy of the symbolic labels speech is given by

$$H(X)=\sum_{i=1}^{40}-\frac{1}{40}\log_2\frac{1}{40}=5.3 \quad \text{bits}$$ (13.41)

Number of bits/sec = Number of symbols/sec × Entropy = 5.3×10 = 53 bps.

Note the above over-simplified calculation ignores the fact that phonemes, like letters of written language, have a non-uniform probability distribution. More importantly it also ignores the constraints imposed on the occurrence of phonemes by dictionary, the rules of grammar and the context of conversation, and it ignores the speaker characteristics and intonation information.

13.8.2.1 The effects of dictionary, grammar and topic on entropy

The occurrence of a speech unit in a correctly constructed speech sentence depends on the previous speech units and is constrained by the dictionary, the grammar, the context and the topic of conversation. The probability of a word in a stream of words should therefore be conditioned on the previous words. This is usually achieved using a language probability in the form of a bigram in which the probability of a word is conditioned on the previous word or a trigram in which the probability of a word is conditioned on the previous two words as explained in Section (). The topic of a conversation can be used to influence the likelihood of a word conditioned on the topic.

Using a conditional probability function, based on an *N*-Gram grammar and a topic *T*, the conditional entropy of words is expressed as

$$H(w_m \mid w_{m-1},\ldots,w_{m-N+1},T) \;=\; -\sum_{i=1}^{M} P_X(w_i \mid w_{m-1},\ldots,w_{m-N+1},T)\log P_X(w_i \mid w_{m-1},\ldots,w_{m-N+1},T)$$

(13.42)

The net effect of conditioning the probability of an observation on the previous observations and on the topic, is to constrain and reduce the randomness and hence the entropy of the observation.

13.8.2.2 The effect of speaker characteristics on speech entropy

One of the main sources of variability and information in speech is the variations of speaker characteristics. Speaker variability is due to the variations in the anatomical characteristics as well as the speaking manner. The variables that convey speaker characteristics are the vocal tract model parameters, pitch value, pitch range, and prosody parameters. These variables naturally increase the dimensionality, the randomness and the entropy of speech.

13.8.3 Probabilistic Speech and Language Models

The probability that a sequence of speech feature vectors X is an acoustic production of a word w may be expressed, using the Bayes' rule, as

$$P(w \mid X) = \frac{f(X \mid w)P(w)}{f(X)}$$

(13.43)

where $P(w|X)$ is the probability of a word w given a sequence of speech observation feature vectors X, $f(X|w)$ is the pdf of the speech observation X

conditioned on the word w and $P(w)$ is the prior probability of w obtained from a language model.

The maximum a posteriori (MAP) estimate of the word w conveyed by the feature sequence X is obtained as the word that maximises $P(w|X)$. Since for a given observation feature sequence X, its probability $P(X)$ is a constant, the MAP estimate may be expressed as

$$\hat{w} = \arg\max_{w} f(X \mid w) P(w) \tag{13.44}$$

A "language" model can be incorporated in the Bayesian probability model of Equation () by conditioning the probability of a word w_m on the sequence of speech observation feature vectors X and on the previous $N–1$ words $w_{m-1}, w_{m-2, \ldots} w_{m-N+1}$ (N-gram grammar) as

$$P(w_m \mid X, w_{m-1}, \ldots w_{m-N+1}) = \frac{f(X \mid w_m) P(w_m \mid w_{m-1}, \ldots, w_{m-N+1})}{f(X)} \tag{13.45}$$

where $P(w_m \mid w_{m-1}, w_{m-2}, \ldots w_{m-N+1})$, the probability word w_m conditioned on the previous $N–1$ words $w_{m-1}, w_{m-2}, \ldots w_{m-N+1}$, is an N-gram language model. The MAP estimate is then given by

$$\hat{w}_m = \arg\max_{w} f(X \mid w_m) P(w_m \mid w_{m-1}, w_{m-2}, \ldots, w_{m-N+1}) \tag{13.46}$$

N-gram language models can be trained using a large database of transcription of speech or simply using texts of books or newspapers. Usually a bi-gram language model $P(w_m \mid w_{m-1})$ in which the occurrence of a word w_m is conditioned on the previous word w_{m-1}, or a tri-gram model $P(w_m \mid w_{m-1}, w_{m-2})$ in which the occurrence of a word w_m is conditioned on the preceding two words w_{m-1} and w_{m-2} are used.

13.8.3.2 Front-end feature extraction

The feature extraction subsystem converts time-domain raw speech samples into a compact and efficient sequence of spectral-temporal feature vectors that retain the phonemic information but discard some of the variations due to speaker variability and noise. The most widely used features for speech recognition are cepstral feature vectors which are obtained from a discrete cosine transform function of the logarithm of magnitude spectrum of speech as described in section xx. The temporal dynamics of speech parameters, i.e. the direction and the rate of change of time-variation of speech features,
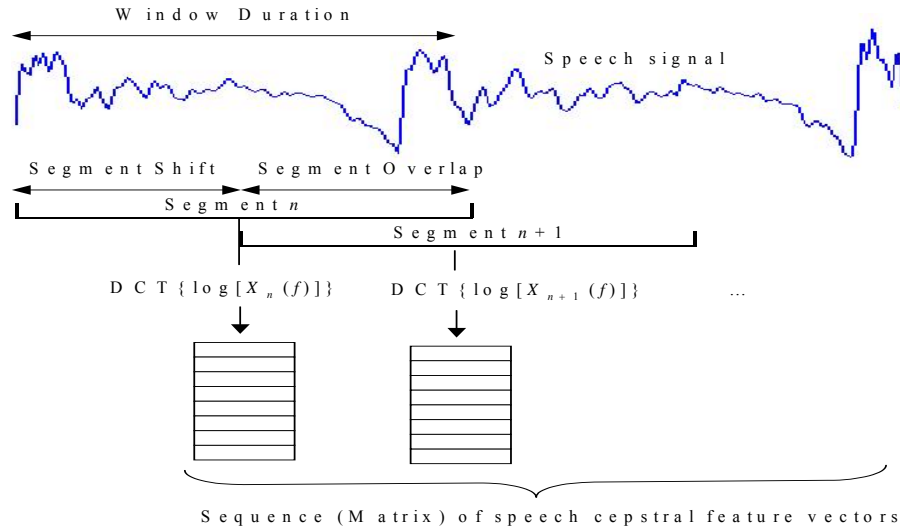
Figure 13.19 – Illustration of Speech Feature Extraction.

play an important rule in improving the accuracy of speech recognition. Temporal dynamics are often modelled by the first and the second order differences of cepstral coefficients as explained in section ().

13.8.3.3 Cepstral Features

For speech recognition, the most commonly used features are cepstral coefficients. Cepstral coefficients are derived from an inverse discrete Fourier transform (IDFT) of logarithm of short-term power spectrum of a speech segment (with a typical segment length of 20-25ms) as:

$$c(m) = \sum_{k=0}^{N-1} \ln[|X(k)|] e^{\frac{j2\pi mk}{N}} \qquad (13.47)$$

where $X(k)$ is the FFT-spectrum of speech $x(m)$. As the spectrum of real-valued speech is symmetric, the DFT in Equation (13.9) can be replaced by a Discrete Cosine Transform (DCT) as:

$$c(m) = DCT\left\{\ln\left[|X(k)|\right]\right\} \qquad (13.48)$$
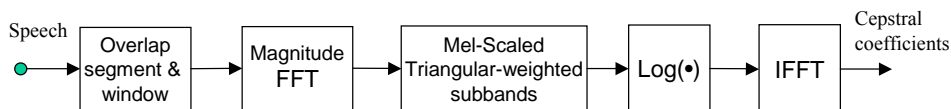

Figure 13.20 Block diagram of a typical cepstral feature extraction system.

As shown in Figure (13.6) *cepstral parameters encode the shape of the log-spectrum*. For example the coefficient $c(0)$ is given by:

$$c(0) = \log |X(0)| + \cdots + \log |X(N-1)| = \log\left[|X(0)| \times \cdots \times |X(N-1)|\right]$$

(13.49)

Note that $c(0)$ is the average of log magnitude spectrum, or equivalently the the geometric mean of magnitude spectrum. Similarly the coefficient $c(1)$ describes the tilt or the slope of the log spectrum which is usually negative for vowels (e.g. '*e*' pronounced as '*iy*') with their energy mostly in low frequencies and positive for consonants (e.g. '*s*') with their energies mostly in higher frequencies. Some useful properties of cepstrum features are as follows:

(a) The lower index cepstral coefficients represent the spectral envelop of speech, whereas the higher indexed coefficients represent fine details (i.e. excitation) of the speech spectrum.

(b) Logarithmic compression of the dynamic range of spectrum, benefiting lower power higher frequency speech components.

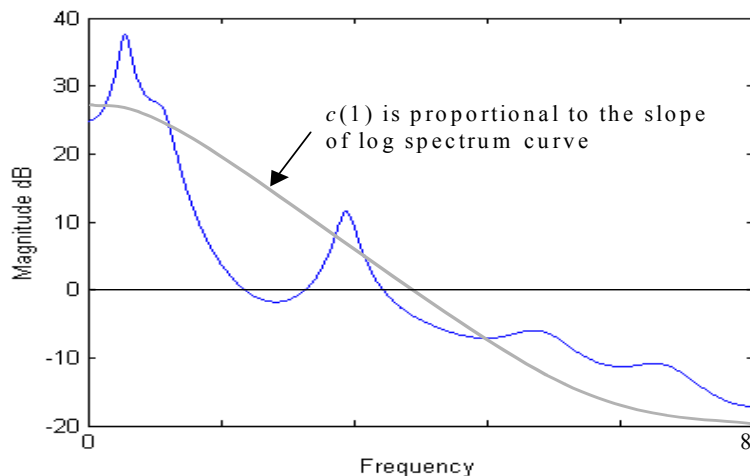(c) Insensitivity to loudness variations of speech if the coefficient $c(0)$ is discarded.

Figure 13.21 Cepstral coefficients encode the shape of the spectrum. For example $c(1)$, corresponding to the DCT basis function shown, encodes the tilt of the spectrum.
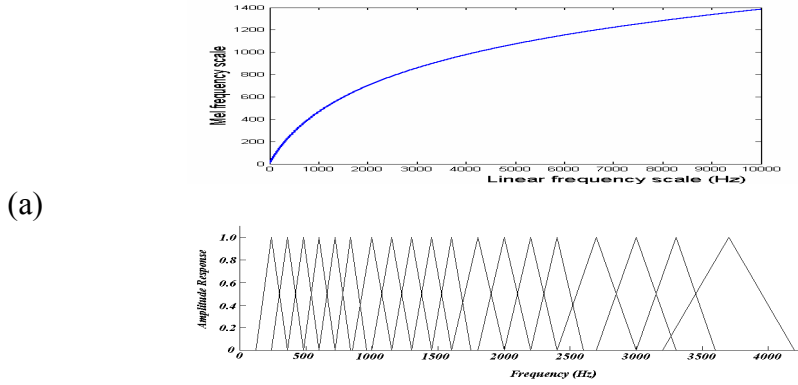
(a)

(b)

Figure 12.22 Illustration of: (a) linear to mel frequency scale mapping, (b) Mel frequency scale bands for a filter bank Analysis of speech.

  (d) As the distribution of the power spectrum is approximately log-normal, the logarithm of power spectrum, and hence the cepstral coefficients, are approximately Gaussian.

  (e) The cepstral coefficients are relatively de-correlated allowing simplified modelling assumptions such as the use of diagonal covariance in modelling the speech feature vectors distribution.

There are a number of variants of cepstral coefficients. The two most common choices for extraction of cepstral coefficients are based on a filter bank model and a linear predictive (LP) model of speech.

13.8.3.4 Mel Frequency Cepstral Coefficients

A widely used form of cepstrum is mel frequency cepstral coefficients (MFCC). To obtain MFCC features, the spectral magnitude of FFT frequency bins are averaged within frequency bands spaced according to the mel scale which is based on a model of human auditory perception. The scale is approximately linear up to about 1000 Hz and approximates the sensitivity of the human ear as:

$$f_{mel} = 1125\log(0.0016f + 1) \qquad (13.50)$$

where f_{mel} is the mel-scaled frequency of the original frequency f in Hz. An example of the mapping process and also the spacing of bands for a 19-channel filter-bank is shown in Figure (13.7).

13.8.3.5 LP-Cepstrum

Cepstrum coefficients can also be derived from linear prediction model parameters. The LP-cepstrum coefficients are given by []

$$c(m) = -a(m) + \sum_{k=1}^{m}(1 - \frac{k}{m})a(k)c(m-k) \qquad 1 \le m \le P \qquad (13.51)$$

where P is the order of the LPC analysis.

13.8.3.6 Temporal Dynamic Features of Speech

Although over a short time segment of say 20 ms speech parameters are assumed to be time-invariant, over a longer segment the parameters of speech vary with time. Furthermore, the time variations of speech convey essential linguistic, prosodic and speaker-characteristic information. Hence, speech parameters, such as linear prediction model coefficients, pitch and excitation gain, are time-varying variables. Temporal difference features are a simple and effective means for description of the trajectory of speech parameters in time. The modelling of difference features is essential in improving the accuracy in speech recognition or the perceptual quality in speech synthesis. For speech recognition relatively simple cepstral difference features are defined by

$$\partial c(m) = c(m+1) - c(m-1) \qquad (13.52)$$

$$\partial\partial c(m) = \partial c(m+1) - \partial c(m-1) \qquad (13.53)$$

where $\partial c(m)$ is the first order time-difference of cepstral features, in speech processing it is also referred as the velocity features. The second order time-difference of cepstral features $\partial\partial c(m)$ is also referred to as the acceleration. Equation (13.14) can be derived from the slope of the minimum mean squared line model of three cepstral samples $c(m-1)$, $c(m)$ and $c(m+1)$.

Dynamic or difference features are effective in improving speech recognition. The use of difference features is also important for improving quality in speech coding, synthesis and enhancement applications.

13.8.4 Statistical Models of Acoustic Features of Speech

For speech recognition, an efficient set of acoustic models is needed to capture the mean and variance of the spectral-temporal trajectory of speech

sounds and to discriminate between different speech sounds. In selecting a speech model we have the following broad options:

(a) Templates of averaged speech feature vector sequences. This is mostly used for isolated-word applications. Time alignment of speech examples of different length and time alignment of a speech example with speech templates is achieved using dynamic time warping method (DTW).

(b) Hidden Markov models (HMMs) or artificial neural networks (ANNs). There are also hybrid models that combine the complimentary power of HMMs and ANNs.

(c) Context-independent models or context-dependent models where several models are used for each word (or phoneme) to capture the variations in the acoustic production of the words caused by the variations of different acoustic context within which words occurs.

For small vocabulary isolated-word recognition, as in name-dialling, a template of cepstral feature vectors can be used to model the acoustic realisation of each word. The template for each word is obtained from one or more examples of the word. For large vocabulary and continuous speech recognition, sub-word models are used due to their efficiency for training and adaptation of models, and for the ease of expansion of the vocabulary size. Hidden Markov models described in chapter 5 are commonly used for medium to large vocabulary speech recognition systems.

13.8.5 Resolutions of Speech Features and Models

The resolution of the features and mopdels employed ina speech recognition system depends on the following factors:

(a) Acoustic feature resolution, i.e. spectral-temporal resolution.

(b) Acoustic model resolution; i.e. the smallest segment modelled.

(c) Statistical model resolution.

(d) Contextual resolution.

Spectral-temporal feature resolution. The spectral and temporal resolution of speech features are determined by the following factors: (i) the speech signal window size for feature extraction (typically 25 ms), (ii) the rate at which speech features are sampled (usually every 5 to 10 ms), and

(iii) speech feature vector dimensions; typically 13 cepstral features plus 13 first difference and 13 second difference cepstral features.

Statistical model resolution. Model resolution is determined by (i) the number of models, (ii) the number of states per model, and (iii) the number of sub-state models per state. For example, when using hidden Markov models, each HMM has N states (typically N=3-5), and the distribution of feature vectors within each state is modelled by a mixture of M multi-variate Gaussian densities. Therefore the model for each phoneme has $N \times M$ Gaussian distributions with each Gaussian density parameterised by a mean feature vector and a covariance matrix. Model resolution also depends on whether full-covariance or diagonal-covariance matrices are used for the Gaussian distribution of speech feature vectors. Using full covariance matrix Gaussian pdfs, the number of Gaussian model parameters for N states, with M mixtures per state and P-dimensional features is $N \times M \times (P^2 + P)$, with typical numbers used (i.e. N=3, M=15, and P=3×13=39 including the first and second difference features) we have $3 \times 15 \times (39^2 + 39)$=70200 parameters. The use of diagonal covariance matrices reduces this to about 3510 parameters.

Model context resolution. The acoustic production of a speech unit is affected by its acoustic context; that is by the preceding and succeeding speech units. In phoneme-based systems, context-dependent triphones (see section xx) are used. Since there are about 40 phonemes, the total number of triphones is 40^3=64000, although many of these cannot occur due to linguistic constraints. The states of triphone models are often clustered and tied to reduce the total number of parameters and hence obtain a compromise between contextual resolution and the number of model parameters used.

13.8.6 Voice-Activated Name Dialling

This section describes a voice-activated name dialling method for mobile phones. The dialling method is developed with the objective of optimising the requirements for high accuracy, simplicity of use and minimum of; memory, computational and power requirement. The system stores an acoustic template for each name along with its text form and telephone number. The text for each name and telephone number are input manually. The acoustic template for each name is formed from a single spoken example of the name. The recognition system is an isolated-word recognition that makes use of the LPC parameters and the voice activity

detector (VAD) of the mobile phone system. LPC-Cepstrum parameters are used as speech features, and the VAD output is used for endpoint detection. The distance metric used for selection of the most likely spoken name is the minimum mean squared distance. In order to minimise the cost of making wrong calls due to speech recognition errors, the best score is compared to a confidence threshold level; if the score is less than the threshold the system prompts: 'Name not Recognised'. The system was evaluated on a Laptop PC. Initial experiments obtained very high accuracy and reliability of 97%. This is not surprising for a speaker-dependent small-vocabulary isolated-word recognition task.

13.8.6.1 Name-Dialling Vocabulary

For name dialling task the vocabulary is simple and consists of N names

$$\text{Names} = \text{sil} <\text{Name1} \mid \text{Name2} \mid \text{Name3} \mid \ldots \mid \text{Name } N> \text{sil}$$

where N is typically 8, and sil indicates silence/noise i.e. no speech activity and the bar '|' indicates 'or' (i.e. Name1 or Name2 or …). The names, numbers and the acoustic templates for each phonebook entry are stored as shown.

Name	Number	Acoustic Template
Luke	08890552077	
William	01763 226226	….
Colin	….	….
Kate	….	….

Table13.4- Illustration of stored codebook of name, number and feature template for name-dialling.

13.8.6.2 Speech Features and Models for Name Dialling

Since mobile phones use LPC models for speech coding, LPC-based cepstrum coefficients and differential cepstral coefficients are used as speech features. An important point is the dimensionality of cepstral features and the feature vector sampling rate for representation of each word. The feature template size is chosen as a compromise between the requirements for high accuracy and low memory storage and computational complexity.
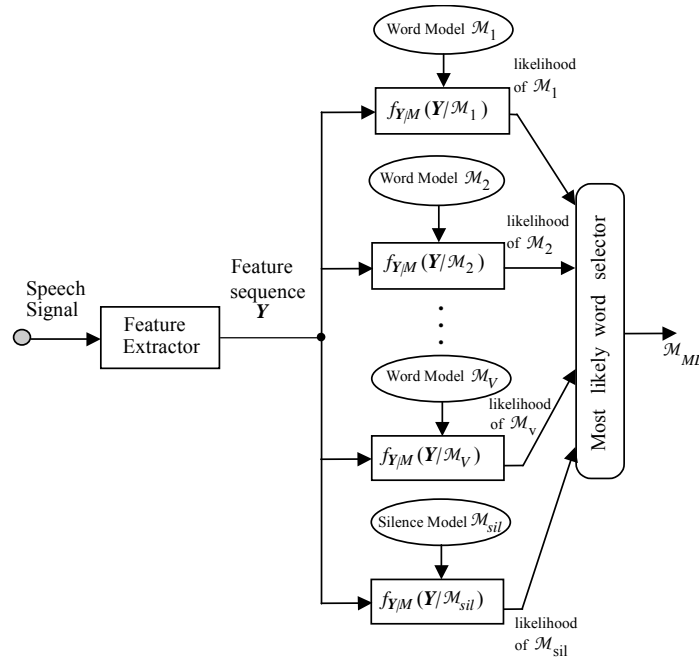
Figure 13.23 Illustration of a speech recognition system. Note that for a simple voice-activated dialling, each word model is a single spoken example of a name; the feature extraction includes duration normalisation; and instead of a probability metric a mean squared distance metric is used.

Hidden Markov models are normally used for speech recognition. However for this relatively simple voice dialling application we simply store and use the cepstral vector sequence from one example of each word as the template for that word.

13.8.6.3 Dealing with Speech Duration Variability

For continuous speech recognition the variability in speech duration is usually dealt with using the model flexibility afforded by the state transitions of a Hidden Markov model together with the Viterbi time warping algorithm. However, for the isolated-word voice-dialling system, there are two relatively simple methods that can be used:

(a) Time-normalisation can be achieved by adjusting the frame shift rate (Figure 13.14) for calculation of cepstrum feature vectors sequence of each input name, to yield a pre-specified number of uniformly-spaced cepstrum-feature vectors across the duration of

the spoken name. For small-vocabulary isolated-word recognition this method works very well indeed.

(b) Time normalisation between two feature vector sequences of different length can be achieved using the dynamic time warping as described next.

13.8.6.4 Time-Alignment: Dynamic Time Warping (DTW)

Speech is a time-varying process in which the duration of a word and its subwords varies randomly. Hence a method is required to find the best time-alignment between a sequence of vector features representing a spoken word and the model candidates. The best time-alignment between two sequences of vectors may be defined as the alignment with the minimum Euclidean distance. For isolated-word recognition the time-alignment method used is dynamic-time warping (DTW) illustrated in figure 13.15. The best matching template is the one with the lowest distance path aligning the input pattern to the template.

For illustration of DTW consider a point (i,j) in the time-time matrix (where i indexes the input pattern frame, and j the template frame) of Figure(13.16), then previous point must have been $(i-1,j-1)$, $(i-1,j)$ or $(i,j-1)$. The key idea in dynamic programming is that at point (i,j) we continue with the lowest accumulated distance path from $(i-1,j-1)$, $(i-1,j)$ or $(i,j-1)$. The DTW algorithm operates in a time-synchronous manner: each column of the time-time matrix is considered in succession (equivalent to processing the input frame-by-frame) so that, for a template of length N, the maximum number of paths considered at any time is N. If $D(i,j)$ is the global distance up to (i,j) and the local distance at (i,j) is $d(i,j)$ then we have the recursive relation

$$D(i, j) = \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] + d(i, j) \quad (13.54)$$

Given that $D(1,1) = d(1,1)$, we have the basis for an efficient recursive algorithm for computing $D(i,j)$. The final global distance $D(n,N)$ gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

13.8.6.5 Distance Metric Score for Word Recognition

The distance metric used for the selection of the nearest name in the memory $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\}$ to the spoken input name X can be the mean squared error or alternatively the mean absolute value of error. The labelling of the input feature matrix X is achieved as

$$Label(X) = \arg\min_k \left\{ |X - \mathcal{M}_k|^2 \right\}$$

$$k = 1, \ldots, N \quad (13.55)$$

where the squared magnitude function is given by

$$|X - \mathcal{M}_k|^2 = \sum_{t=1}^{T} \sum_{i=1}^{P} |X(i,t) - \mathcal{M}_k(i,t)|^2$$
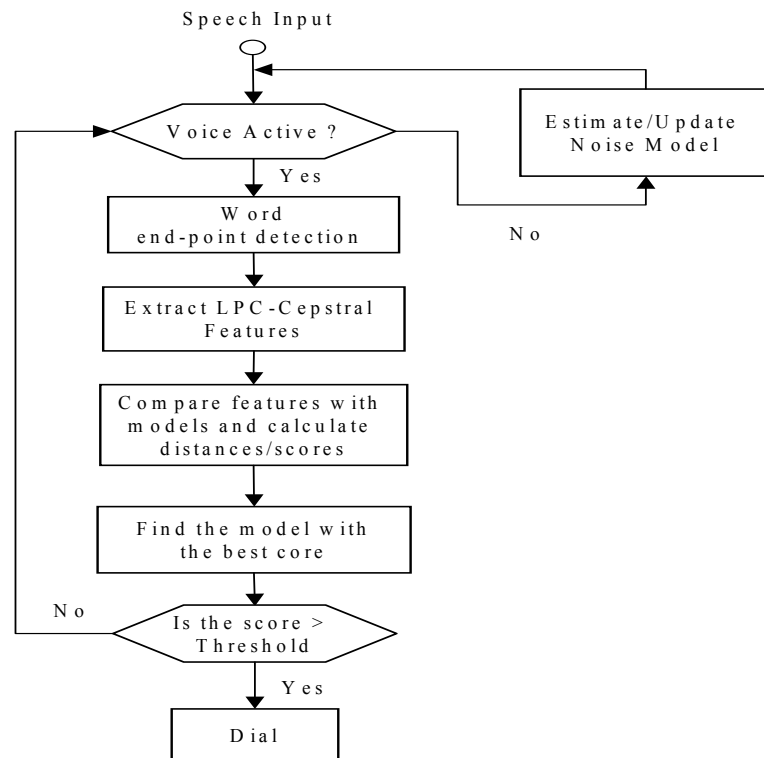
$$(13.56)$$

Figure 13.24 Illustration of voice-activated dialling.

Minimising Dialling Error: Calculating a Confidence Threshold

Recognition errors in voice-activated dialling can result in wrong phone calls. To minimise recognition errors the speech recognition scores are analysed and a confidence threshold is required so that when the metric distances between the spoken word and the best and second best candidates are too close then the system will output a message such as *"Name Not Recognised"*, The confidence threshold depends on the metric and can set experimentally to achieve a desired risk/cost ratio.

Automatic Activation of Name Dialling

The name dialling can be activated by either pressing a button on the phone or by automatic speech/noise detection. If automatic speech/noise detection is used then a fairly high threshold level should be chosen to minimise the risk inadvertent name dialling. Name dialling for mobile phone has a real practical value in hands-free environment, for example when the user is driving etc.
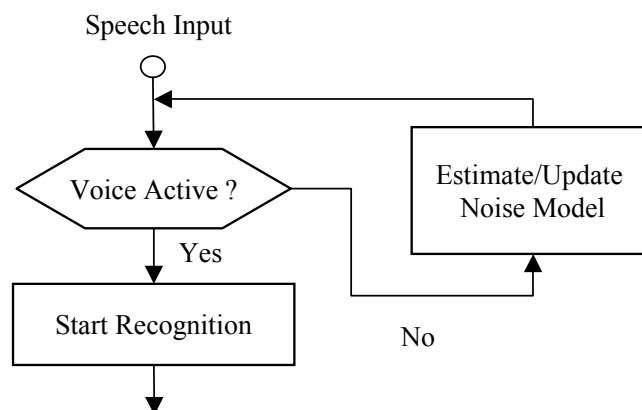
Speech Input

Voice Active ?

Yes

Start Recognition

Estimate/Update Noise Model

No

Figure 13.25 Illustration of automatic activation of name dialling.

References

1. RABINER L.R. and JUANG B.H. (1993) Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

2. L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.*

Prentice Hall, Englewood Cliffs, Newb Jersey, 1978.

3. Furui S. (1989), Digital Speech Processing, Synthesis and Recognition, Marcel Dekker.

4. John R. Deller, John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Maxwell Macmillan International, 1993.

5. Emmanuel C. Ifeachor and Barrie W. Jervis. *Digital Signal Processing: A practical approach*. Addison-Wesley, 1993.

6. B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, London, second edition, 1982.

7. M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP*, pages 208-211, 1979.

8. J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561-580, April 1975.

9. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738-1752, 1990.

10. Griffin D. W., Lim J.S. "Multiband-excitation vocoder" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-36(2) pp.236-243