to add two probabilities, one exponential and one logarithmic operation are usually required. These operations can be avoided by using a method which allows the numbers to be added while in their logarithmic form[5]. Considering step 4 in the sequence of calculations described above, both the exponential and the logarithmic operation can be avoided by using a pre-computed look-up table to store the values of $\log(1 + B/A)$ in terms of $\log(B/A)$. Thus steps 3 and 4 can be replaced by a single table look-up operation, with $\log(B/A)$ as input (i.e. the value already stored in C at step 2). The output is $\log(1 + B/A)$, which can again be stored as the new value of C. Moderate accuracy in the value of $\log(A + B)$ can be achieved with a small look-up table. For example, a 1% accuracy for $A + B$ enables values of B/A of less than 0·01 to be ignored, and the look-up table for the larger values of B/A only needs entries for 115 equally spaced values of $\log(B/A)$.

If the above method is implemented using a suitable scale factor for the logarithms, it is then even possible to make all the probability calculations for recognition and parameter estimation using integer arithmetic on logarithmically coded numbers. No multiplications would be required with the VQ method, and no exponential functions would be needed when using Gaussian distributions. The 1% error proposed above should have very little effect on the re-estimation, but the error could easily be reduced if necessary by using a larger look-up table.

9.13 RELATIONSHIP BETWEEN DTW AND A SIMPLE HMM

It is interesting to compare the Markov probability calculation with the cumulative distance formula for a simple asymmetric dynamic programming algorithm in which each input frame occurs exactly once in the distance calculation. If the DP uses a squared Euclidean distance metric, the recognition process can be regarded as a special case of HMM Viterbi decoding, in which the word models have one state per template frame, and the features are assumed to be normally distributed with unit variance.

To clarify this relationship, we will return to the Viterbi calculation using logarithms of probabilities, given in Equation (9.52). The value of $b_j^L(y_t)$ according to an uncorrelated normal distribution is given by Equation (9.53), where we are now assuming that $\sigma_{jk} = 1$ for all states j and for all features k. Hence

$$\sum_{k=1}^{K} \log(\sigma_{jk}) = 0 ,$$

and recursive calculation of $\hat{\alpha}_j^L(t)$ simplifies to:

$$\hat{\alpha}_j^L(t) = \max_{\text{over } i} \left(\hat{\alpha}_i^L(t-1) + a_{ij}^L \right) - \frac{K}{2} \log(2\pi) - \frac{1}{2} \sum_{k=1}^{K} \left(y_{kt} - \mu_{jk} \right)^2 . \qquad (9.55)$$

The term $K/2 \log(2\pi)$ is a constant, which will scale the likelihood calculation but will not affect the choice of optimal state sequence. Thus the only quantities that need be considered at each frame are the logarithms of the transition probabilities

[5] This method was described by Kingsbury and Rayner (1971) for a completely different application.

Ex.1019 / Page 58 of 102
AMAZON.COM, INC.

and the square of the Euclidean distance between the observed features at time t and the means for model state j.

As maximizing $\hat{\alpha}_j^L(t)$ is equivalent to minimizing $-\hat{\alpha}_j^L(t)$, this recognition task can be regarded as one of minimizing a distance comprising $-a_{ij}^L$, the negative logarithm of the transition probability (which must itself be positive), *plus* the squared Euclidean distance of the features from their model mean values. Thus we have a simple DP algorithm in which the $-a_{ij}^L$ terms are interpreted as timescale distortion penalties. Where only slopes of 0, 1, and 2 are permitted, as is the case for the HMM in Figure 9.1, the time distortion penalties for other values of slope are –log(0), and are therefore infinite.

9.14 STATE DURATIONAL CHARACTERISTICS OF HMMS

The probability of a model staying in the same state, i, for successive frames is determined only by the transition probability, a_{ii}. The expected number of frames it will stay in state i is $1/(1-a_{ii})$, so a value of $a_{ii} = 0{\cdot}9$ would be suitable for using one state to model, for example, a steady fricative sound whose expected duration is 10 frames. Although the expected total duration in state i in this case is 10 frames, the most likely duration is only one frame, with a probability of 0·1. The probabilities for longer durations decrease exponentially, as shown in trace (i) of Figure 9.2(b). This distribution is often referred to as a **geometric distribution** because the probabilities for successive numbers of frames form a geometric progression.

For any state representing a particular phonetic event, this type of duration distribution is obviously not sensible. For any such event there will be a most probable duration, with reducing probability for both shorter and longer durations. If many more states are available, the durational characteristics of the model can be improved, but only if a long steady region is modelled by a sequence of states with very similar feature p.d.f.s and the total likelihood method is used to calculate the word probability. For example, consider a group of four identical states with a repeat probability of 0·6, as shown in trace (ii) of Figure 9.2. The expected duration for the group is 10 frames, as it is for the single state shown in trace (i). However, in the case of the group of states, the variation of probability with duration is much more appropriate for speech sounds within a word. This more realistic distribution arises because, while there is only one possible way of going through the states in the minimum number of frames, there are more possible paths for longer frame sequences. However, the improved shape of duration distribution given by this state-splitting approach relies on using total likelihood probability calculations, whereas the Viterbi algorithm is generally used for recognition.

A simple method which can be used with the Viterbi algorithm involves merely imposing a minimum and a maximum duration on state occupancy. Such duration constraints can be achieved with an easy modification to the recognition algorithm, and can give worthwhile performance benefits. Many other methods have been proposed for improving the duration characteristics of HMMs, including some that model duration distributions of each state explicitly. These methods generally give greater benefits than simple duration constraints, but at the expense of more computation and some increase in the number of model parameters.

Ex.1019 / Page 59 of 102
AMAZON.COM, INC.

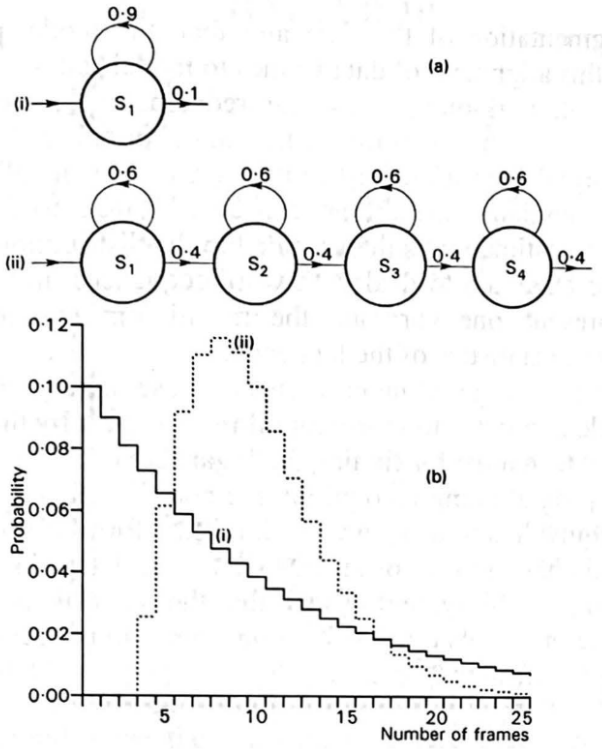


Figure 9.2 **(a)** Two arrangements of states, each with an expected occupation time of 10 frames. **(b)** Probability of occupancy of groups of states in the model sections shown in (a).

CHAPTER 9 SUMMARY

- The performance of pattern-matching speech recognizers is improved by representing typical characteristics of speech patterns in a way that also takes account of variability, which can be achieved by using a stochastic model of each word. Hidden Markov models (HMMs) represent each word as a sequence of states, with transition probabilities between each state and its permitted successors, and probability distributions defining the expected observed features for each state. A recursive formula can be used to calculate the probability that each word model will produce the observed data. The model with the highest probability is assumed to represent the correct word.
- Computation can be saved by using the Viterbi dynamic programming algorithm to calculate the probability of producing the data from only the most likely path through the states. This probability will always be less than the true probability, but the effect on recognition performance is usually very small and the Viterbi algorithm is generally adopted for HMM recognition.
- For each word model the transition probabilities and the probability distributions of the feature vectors can be found by the Baum–Welch re-estimation process. This process iteratively refines initial guesses to improve the model's representation of a set of training examples of the word, taking into account all possible paths through the states of the model.
- An alternative approach to estimating model parameters is to use a Viterbi training procedure, in which the initial guesses are used to find the most likely

Ex.1019 / Page 60 of 102
AMAZON.COM, INC.

state-level segmentation of the data and then the model parameters are re-estimated for this alignment of data frames to model states.

- Vector quantization is one method for reducing the set of possible feature vectors to a number for which robust training is possible. A parametric model, such as the normal (Gaussian) distribution or, more generally, a weighted sum (mixture) of Gaussian distributions, can also be used to describe the feature statistics. The re-estimation is then applied to the distribution parameters.
- HMMs can be extended to deal with word sequences, in which each state of the model represents one word, and the transition probabilities are determined by word sequence statistics of the language.
- One way of overcoming scaling problems because of very small numbers in the probability calculations is to represent all the numbers by their logarithms, and to use a special technique for finding the logarithm of the sum of two numbers.
- The dynamic programming recognition method described in Chapter 8 can be shown to be equivalent to using a very simplified form of HMM.
- The durational characteristic of an HMM state is determined only by the self-loop transition probability, and is such that the most likely duration is always only one frame and probabilities for longer durations decrease exponentially, so forming a geometric progression.

CHAPTER 9 EXERCISES

E9.1 What is the significance of the word 'hidden' in hidden Markov models?

E9.2 Why is it not necessary to explicitly consider all possible state sequences when calculating the probability of an HMM generating observed data?

E9.3 What is the essential difference between the Viterbi algorithm and the total likelihood method when calculating the probability of a word model generating observed data? What practical advantages can be gained by using the Viterbi algorithm for recognition?

E9.4 How can the form of an HMM be constrained by choice of initial parameters provided for re-estimation?

E9.5 What is the purpose of the 'vector quantization' sometimes used in HMMs?

E9.6 What are the benefits of using normal distributions to model feature statistics for HMMs? What are the limitations of simple normal distributions and how can these be overcome?

E9.7 How do the calculations required for Viterbi training differ from those for Baum–Welch re-estimation?

E9.8 What are the practical difficulties associated with implementing forward and backward probability calculations? What solutions are usually adopted?

E9.9 How can a simple HMM be interpreted as equivalent to a DTW recognizer?

E9.10 Why are the state durational characteristics of HMMs not very appropriate for modelling speech? What are the effects on duration characteristics if a single state is replaced by a sequence of several identical states?

Ex.1019 / Page 61 of 102
AMAZON.COM, INC.

CHAPTER 10

Introduction to Front-end Analysis for Automatic Speech Recognition

10.1 INTRODUCTION

The term "front-end analysis" refers to the first stage of ASR, whereby the input acoustic signal is converted to a sequence of acoustic **feature vectors**. As explained in Section 8.3, the short-term spectrum provides a convenient way of capturing the acoustic consequences of phonetic events. Ideally the method of front-end analysis should preserve all the perceptually important information for making phonetic distinctions, while not being sensitive to acoustic variations that are irrelevant phonetically. As a general policy for ASR, it seems desirable not to use features of the acoustic signal that are not used by human listeners, even if they are reliably present in human productions, because they may be distorted by the acoustic environment or electrical transmission path without causing the perceived speech quality to be impaired. Over the years many different front-ends have been tried, for use first with DTW recognizers and, more recently, with HMM systems. These front-ends vary in the extent to which they incorporate knowledge about human auditory perception, but currently the most successful analysis methods include at least some of the known properties of perception. These successful methods are, however, also characterized by a compatibility with the mathematical techniques that are generally used in HMM recognizers (as will be explained later). In this chapter we will introduce various aspects of front-end analysis for ASR.

10.2 PRE-EMPHASIS

The spectrum of voiced speech is characterized by a downward trend, whereby frequencies in the upper part of the spectrum are attenuated at about 6 dB/octave. This downward trend is due to a combination of the typical –12 dB/octave slope of the glottal source spectrum with the +6 dB/octave lift given by the radiation effect due to the lips (see Chapter 2). For the purpose of front-end analysis, it is common to compensate by applying a **pre-emphasis** of 6 dB/octave so that the analysed signal has a roughly flat spectral trend. This pre-emphasis is easily applied to the speech signal as the first processing stage. Although the above argument for pre-emphasis only applies to voiced regions, in practice it is usually applied throughout without causing any obvious problems for the analysis of voiceless regions.

10.3 FRAMES AND WINDOWING

Due to physical constraints, the vocal tract shape generally changes fairly slowly with time and tends to be fairly constant over short intervals (around 10–20 ms). A

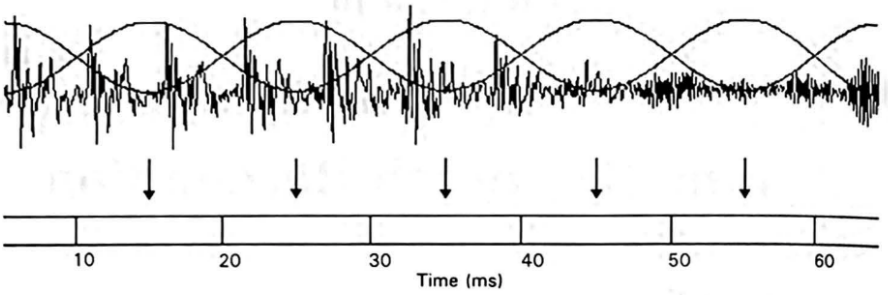


Figure 10.1 Analysis of a speech signal into a sequence of frames. This example shows a 20 ms Hanning window applied at 10 ms intervals to give a frame rate of 100 frames/s.

reasonable approximation is therefore to analyse the speech signal into a sequence of **frames**, where each frame is represented by a single feature vector describing the average spectrum for a short time interval.

Prior to any frequency analysis, each section of signal is multiplied by a tapered **window** (usually a **Hamming** or **Hanning** window). This type of windowing is necessary to reduce any discontinuities at the edges of the selected region, which would otherwise cause problems for the subsequent frequency analysis by introducing spurious high-frequency components into the spectrum. The length of each analysis window must be short enough to give the required time resolution, but on the other hand it cannot be too short if it is to provide adequate frequency resolution. In addition, because the analysis is normally performed at a fixed time interval, during voiced speech the window must be long enough so that it is not sensitive to exact position relative to the glottal cycle (i.e. there needs to always be at least one complete glottal cycle in the main part of the window). Long windows also have the advantage of smoothing out some of the random temporal variation that occurs in unvoiced sounds such as fricatives, but at the expense of blurring rapid events such as the releases of stop consonants. A common compromise is to use a 20–25 ms window applied at 10 ms intervals (giving a **frame rate** of 100 frames/s and an overlap between adjacent windows of about 50%), as shown in Figure 10.1.

10.4 FILTER BANKS, FOURIER ANALYSIS AND THE MEL SCALE

In Section 8.3 we introduced a speech signal representation using a filter bank with channels whose bandwidth and spacing increase with frequency (motivated by psychophysical studies of the frequency resolving power of the human ear). A convenient implementation of filter-bank analysis involves applying a Fourier transform. The output of the Fourier analysis will usually be at a finer frequency resolution than is required, especially at high frequencies. Thus the Fourier magnitudes are summed into a smaller number of channels, whose bandwidth and spacing conform to a perceptual scale such as the Bark or mel scale (see Section 3.5). Typically no more than 20 such channels are used for speech with a 4 kHz bandwidth, with a few additional channels being needed for higher-bandwidth signals. As already explained in Section 8.3, it is advantageous for the filter-bank output to represent power logarithmically, which reflects the phonetic

Ex.1019 / Page 63 of 102
AMAZON.COM, INC.

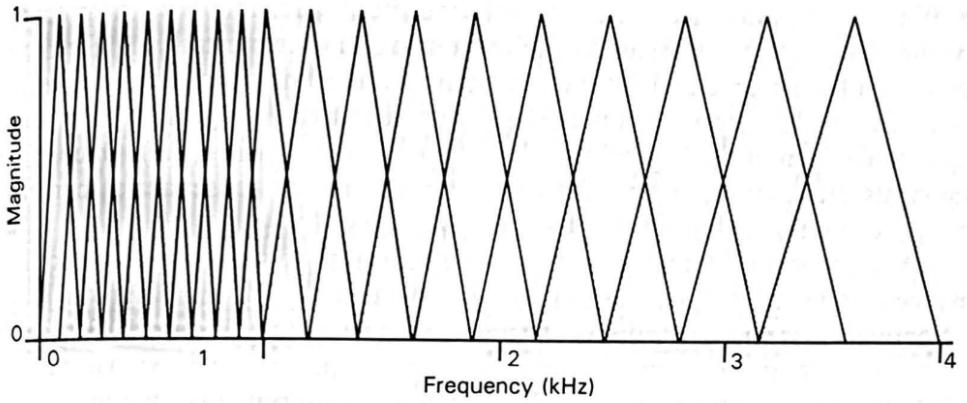


Figure 10.2 Triangular filters of the type suggested by Davis and Mermelstein (1980) for transforming the output of a Fourier transform onto a mel scale in both bandwidth and spacing.

significance of level variations and accords with evidence of a similar compressive non-linearity in auditory systems (see Chapter 3). A consequence of the logarithmic compression is that, when sampled from representative speech over a long period of time, the distribution of the energy in each of the channels tends to follow a Gaussian distribution, and is therefore compatible with any Gaussian assumptions that are made in the modelling.

Figure 10.2 shows a set of triangular 'filters' that can be used to compute a weighted sum of Fourier spectral components, so that the output of the process approximates to a mel scale. Here the centre frequencies of the filters are spaced equally (at intervals of 100 Hz) on a linear scale from 100 Hz to 1 kHz, and equally on a logarithmic scale above 1 kHz. (Other slightly different spacings are also often used.) Each filter's magnitude frequency response is triangular in shape, and is equal to unity at the centre frequency and decreases linearly to zero at the centre frequencies of the two adjacent filters. This configuration of mel filters, which is now very widely used in ASR, was suggested by Davis and Mermelstein (1980).

One option is to use the output of a filter-bank analysis to provide the recognition features directly. However, although filter-bank energies were widely used and achieved a fair amount of success as acoustic features in early recognition systems, there are substantial advantages to be gained by applying further transformations and this approach is the more usual choice nowadays.

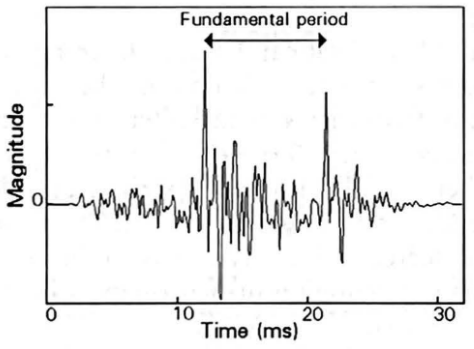
10.5 CEPSTRAL ANALYSIS

The frequency resolution that is given by Fourier analysis applied to a 20–25 ms window of speech is generally sufficient to resolve the individual harmonics of the voiced excitation source, as well as showing the spectral shaping that is due to the vocal tract. Because the filtering operation of the vocal tract is the most influential factor in determining phonetic properties of speech sounds, it is desirable to separate out the excitation component from the filter component. The vocoders described in Chapter 4 are based on this principle. **Cepstral analysis** is another technique for estimating a separation of the source and filter components. Here the starting point is the observation that passing an excitation signal through a vocal-

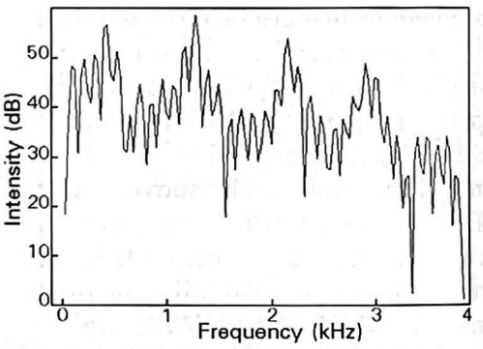
Ex.1019 / Page 64 of 102
AMAZON.COM, INC.

tract filter to generate a speech signal can be represented as a process of **convolution** in the time domain, which is equivalent to multiplying the spectral magnitudes of the source and filter components. When the spectrum is represented logarithmically, these components are *additive*, because the logarithm of a product is equal to the sum of the logarithms ($\log(A \times B) = \log(A) + \log(B)$). Once the two components are additive, it is relatively straightforward to separate them using filtering techniques.

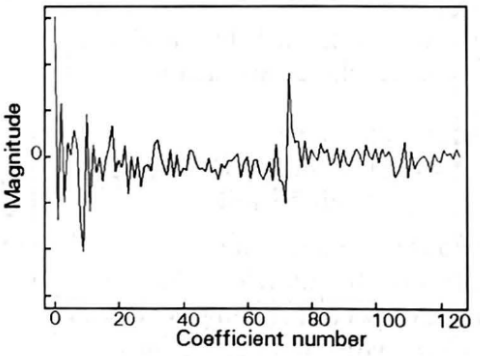
A typical logarithmic spectrum cross-section shows the rapidly oscillating component due to the excitation superimposed on a more gradual trend representing the influence of the vocal tract resonances (see Figure 10.3(b)). If we now imagine that this combined shape represents a time-domain signal, the rapid oscillations would correspond to high-frequency components, while the more gradual changes would be due to low-frequency components. If a Fourier transform were applied, the two components would therefore appear at opposite ends of the resulting spectrum. Thus by starting with the log magnitude spectrum and computing a Fourier transform, to obtain the so-called **cepstrum** (an anagram of "spectrum"), the excitation is effectively separated from the vocal-tract filtering, as shown in Figure 10.3(c). In fact, because the log magnitude spectrum is a symmetric function, the Fourier transform can be conveniently simplified to a



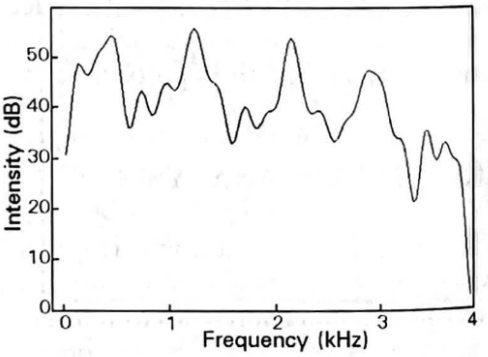
(a) Windowed speech waveform (32 ms at 8 kHz sampling rate).



(b) Log spectrum (from a Fourier transform).



(c) Cepstrum computed from the log spectrum shown in (b).



(d) Log spectrum reconstructed from the first 40 cepstral coefficients in (c).

Figure 10.3 Analysing a section of speech waveform to obtain the cepstrum and then to reconstruct a cepstrally smoothed spectrum.

Ex.1019 / Page 65 of 102
AMAZON.COM, INC.

discrete **cosine transform (DCT)**. For a spectral representation comprising N channels with log magnitudes A_1 to A_N, the DCT can be computed as follows:

$$c_j = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} A_i \cos\left(\frac{\pi\, j(i-0.5)}{N}\right) \quad \text{for } 0 \le j < N, \tag{10.1}$$

where c_j is the j^{th} **cepstral coefficient**. When $j = 0$, Equation (10.1) simplifies so that c_0 is proportional to the mean of the individual log channel signals A_1 to A_N. The c_1 term reflects the balance between energy at low frequencies and energy at high frequencies. As j increases, c_j captures increasingly fine spectral detail: first overall spectrum shape, then general formant structure, then more detailed spectral structure between the formants and, at high values of j, the excitation structure. There is no simple relationship between the c_j terms and the formants. However, for periodic speech the effect of the excitation source tends to be seen as a clear 'spike' at the pitch period duration (see Figure 10.3(c)). Cepstral analysis is therefore one method that can be used to estimate fundamental frequency. For example, in Figure 10.3(c) the spike occurs at c_{73} which, for the sampling frequency of 8 kHz (i.e. a sample duration of 0.125 ms), corresponds to a fundamental period of $73 \times 0.125 = 9.125$ ms. This value can be seen to be roughly equal to the interval between the pitch pulses in Figure 10.3(a).

Although the cosine transformation given by Equation (10.1) ensures that the Euclidean distance in transformed space is exactly equal to the distance between the sets of untransformed channel signals, the information that is of phonetic significance becomes concentrated in the lower-order terms. Filtering the cepstrum (a process usually referred to as **liftering**) can be applied to remove certain components or alter the relative influence of the different components. A simple lifter is one which simply truncates the cepstral sequence, by giving a weight of one to the low coefficients (up to some specified index) and a weight of zero to all the higher coefficients. By setting the cut-off point to just below the coefficient corresponding to the pitch period, most of the influence of the fundamental period is effectively removed from the spectrum. This process is shown in Figure 10.3(d), in which the spectrum has been re-constructed (by a Fourier transform) from just the low-order cepstral coefficients. The resulting spectrum can be seen to be much smoother than the original and show the formant peaks more clearly. The lower the cut-off point is set, the more detail will be removed and the smoother the spectrum will be. The process of smoothing the spectrum by truncating the sequence of cepstral coefficients is often referred to as **cepstral smoothing**.

The effectiveness of cepstral analysis for separating out the fundamental-frequency component of a speech signal depends on the frequency of the fundamental relative to the frequencies of the formants. The method generally works best for adult male speech (as shown in the example in Figure 10.3). For typical female and children's speech both the pitch and formant frequencies are higher, but the pitch increases more relative to the formant frequencies and so the cepstrum gives a less clear separation of the excitation component. It is therefore more difficult to set a cut-off point for cepstral smoothing that removes the pitch influence without also removing useful information about the formant structure.

In addition to the beneficial effect of concentrating on the information that is of greatest phonetic significance, discarding the high-order cepstral coefficients

Ex.1019 / Page 66 of 102
AMAZON.COM, INC.

reduces the number of features, so less computation is needed for the pattern-matching process.

The cepstrum has another desirable property for use in speech recognition. For typical speech signals it is found that, in contrast with the original channel signals, the variation of the separate coefficients tends to be uncorrelated. As a consequence, when using HMMs with continuous-density probability distributions, full covariance matrices can be replaced by much simpler diagonal covariance matrices (see Section 9.8.1) without any great loss in performance. Using diagonal covariance matrices substantially reduces both the computational requirements and the number of parameters needed to represent each distribution.

Cepstral coefficients have the property that (ignoring coefficients that are associated with pitch) both the variance and average numerical values decrease as the coefficient index increases (see Figure 10.3(c)). A consequence for a DTW recognizer using a simple Euclidean distance metric is that the distance calculation is affected most by the lowest-order coefficients and the coefficients that are more related to formant structure tend to be given insufficient weight. A solution that has often been adopted is to apply a lifter with a weighting for each coefficient that acts to roughly equalize the variances for the different coefficients. The problem does not arise when using probability distributions in HMM systems, because the variance is accommodated in the probability calculations. The liftering is often still applied, however, because the effect of making the variances of all the features cover a similar range makes it easier to study model parameters and to place restrictions on variances as part of re-estimation (see Section 11.4.1).

As explained above, the c_0 coefficient is proportional to the mean of the log channel signals and therefore provides an indication of overall level for the speech frame. Sometimes c_0 is included in the feature set, but often it is discarded and replaced by a different energy measure that is derived from the true signal energy. The energy in each frame will depend on overall speaking level, but for identifying sounds the most relevant factor is the *relative* level for different frames in an utterance. Therefore, for those applications for which the whole utterance becomes available before recognition needs to start, the measured energy is often normalized with respect to the maximum energy found over all frames in the utterance. (See Section 8.3.2 for further discussion about measures of speech level.)

In order to retain the advantages of a perceptually motivated filter-bank analysis, for ASR the cosine transform is usually applied to the output of non-linearly spaced filter-bank channels (see Section 10.4 above). A popular choice is to use **mel-frequency cepstral coefficients** (**MFCCs**), which are obtained by applying a DCT to the output of mel filters such as the ones shown in Figure 10.2. An acoustic representation using MFCCs is often simply referred to as a mel **cepstrum**. As explained above, it is generally advantageous to discard the higher-order coefficients. For example, with 8 kHz bandwidth speech, there might be 24 mel channels but only the first 12 MFCCs are generally used in the final feature set. Although the use of the non-linear filter-bank means that the cosine transform no longer gives a simple separation of the excitation from the vocal-tract filtering (and much of the excitation effect will usually have already been smoothed out by the mel averaging), the truncation of the cepstral sequence has a general spectral smoothing effect that is normally desirable because it tends to remove phonetically irrelevant detail.

Ex.1019 / Page 67 of 102
AMAZON.COM, INC.

10.6 ANALYSIS BASED ON LINEAR PREDICTION

An alternative to filter-bank methods for representing the short-term spectrum is to derive linear prediction (LP) coefficients (usually called LPC analysis because of its origin in linear predictive coding, see Chapter 4). In the past, mainly during the 1970s, many recognizers were built using LPC-derived features and these systems generally gave performance comparable with that obtained from recognizers using filter-bank methods. During the 1980s it became more popular to use LPC-derived cepstral coefficients rather than the LP coefficients themselves because, as in the case of the filter-bank representation, the addition of the cepstral transformation was found to improve recognition performance. A convenient method exists for computing cepstral coefficients directly from the LP coefficients. LP analysis has the advantage that it produces an estimate of the smoothed spectrum, with much of the influence of the excitation removed. However, there is less freedom to apply non-linear processing to combat noise than there is with a filter-bank front-end. In addition, LPC inherently gives uniform weighting to low- and high-frequency regions of the spectrum. A non-linear frequency scale can be incorporated, but complicates the analysis to a greater extent than when using filter-bank methods.

Perceptual linear prediction (PLP) (Hermansky, 1990) is one LP-based analysis method that successfully incorporates a non-linear frequency scale and other known properties from the psychophysics of hearing. In PLP analysis, a Fourier transform is first applied to compute the short-term power spectrum, and the perceptual properties are applied while the signal is represented in this filter-bank form. The spectrum is transformed to a Bark scale, and this spectrum is pre-emphasized by a function that approximates the sensitivity of human hearing at different frequencies (see Figure 3.5). The output is compressed to approximate the non-linear relationship between the intensity of a sound and its perceived loudness. The all-pole model of LPC is then used to give a smooth, compact approximation to the simulated auditory spectrum, and finally the LP parameters are usually transformed to cepstral coefficients for use as recognition features. Apart from the use of LPC to achieve spectral smoothing, PLP analysis is very similar to MFCC analysis, but with perceptual properties incorporated in a way that is more directly related to psychophysical results (see Table 10.1 for a comparison of the two methods). In recent years a number of recognition systems have used PLP-based cepstral coefficients as acoustic features, and experimental evidence suggests that overall they give performance that is comparable with that obtained using MFCCs.

Table 10.1 Comparison between the properties of PLP cepstral coefficients and typical MFCCs.

MFCCs	PLP cepstral coefficients
Cepstrum-based spectral smoothing	LPC-based spectral smoothing
6 dB/octave pre-emphasis applied to speech waveform	equal-loudness pre-emphasis applied to spectrum
triangular mel filters	critical-band filters
logarithmic amplitude compression	cube root amplitude compression

Ex.1019 / Page 68 of 102
AMAZON.COM, INC.

10.7 DYNAMIC FEATURES

In the HMM probability calculations (see Section 9.3), the probability of a given acoustic vector corresponding to a given state depends only on the current vector and the current state, and is otherwise independent of the sequence of acoustic vectors preceding and following the current vector and state. It is thus assumed that there is no dependency between the observations, other than through the underlying state sequence. In reality, however, an acoustic feature vector representing part of a speech signal is highly correlated with its neighbours. In fact, it is often the dynamic characteristics of the features that provide most information about phonetic properties of speech sounds (related to, for example, formant transitions or the closures and releases of stop consonants). These correlations can be captured to some extent by augmenting the original set of ('**static**') acoustic features (such as MFCCs) with dynamic features that are a measure of the change in the static features. These dynamic features are often referred to as **time derivatives** or **deltas**. One way of computing the delta features is by simple differencing between the feature values for two frames either side of the current frame:

$$\Delta y_t = y_{t+D} - y_{t-D}, \tag{10.2}$$

where D represents the number of frames to offset either side of the current frame and thus controls the width of the window over which the differencing operation is carried out. Typically D is set to a value of 1 or 2.

Although time-difference features have been used successfully in many systems, they are sensitive to random fluctuations in the original static features and therefore tend to be 'noisy'. A more robust measure of local change is obtained by applying linear regression over a sequence of frames:

$$\Delta y_t = \frac{\sum_{\tau=1}^{D} \tau \left(y_{t+\tau} - y_{t-\tau} \right)}{2 \sum_{\tau=1}^{D} \tau^2} \tag{10.3}$$

With linear regression, a value of $D = 2$ is the usual choice for an analysis frame rate of 100 frames/s. This regression window of five frames (50 ms) is long enough to smooth out random fluctuations, yet short enough to capture local dynamics.

The delta features described above are first-order time derivatives, which can in turn be used to calculate second-order time derivatives (sometimes referred to as **delta-deltas**). Including first-order time derivative features usually gives a large gain in recognition performance, and adding second-order derivatives (which capture changes in the first-order dynamics) tends to give an additional but smaller improvement. The majority of current HMM systems incorporate first-order derivative features, most often applied to a basic feature set of MFCCs and an energy feature, and many also include second-order derivatives. Most of the benefit from derivative features is due to their ability to capture dynamic information. However, these features also have the useful property that they are not affected by any constant or slowly changing disturbances to the signal (such as linear filtering in microphone pre-amplifiers and on telephone channels, for example), provided that these distortions are additive in the feature domain (see Section 11.2).

Ex.1019 / Page 69 of 102
AMAZON.COM, INC.

10.8 CAPTURING THE PERCEPTUALLY RELEVANT INFORMATION

Both in Chapter 3 and at the beginning of the current chapter we explained the desirability of capturing properties of human phonetic perception in the front-end analysis for ASR. Analysis methods such as PLP take into account several known facts about the lower levels of human auditory processing. However, there is no attempt to model higher-level auditory processing or more specific properties of speech perception in any of the analysis methods that have been described above.

It is now well established that the frequencies of the speech formants, particularly the first and second, are vitally important phonetically. Relative formant amplitudes are much less important, and the detailed structure of the lower-level spectral regions between formants is of almost no consequence. There would therefore seem to be potential for better performance in ASR if these factors could be taken into account when designing acoustic analysis methods and distance metrics. Although auditory models have shown considerable promise for incorporating into systems for ASR (see Section 3.7), these types of features have not yet replaced more general spectral features such as MFCCs or PLP-cepstra as the preferred choice in HMM-based systems. It is possible that substantial changes in the design of the recognizers themselves will be required before it will be possible to gain the full benefit from incorporating auditory models. We will return to this issue in Chapter 16, when we will also discuss the prospects and issues for extracting and using formant information more explicitly in ASR.

10.9 GENERAL FEATURE TRANSFORMATIONS

The DCT is one orthogonal transformation that reduces the dimensionality of a filter-bank output by concentrating the most useful information into a small number of features. Other orthogonal transformations for data reduction include **principal components analysis (PCA)** and **linear discriminant analysis (LDA)**. PCA performs a linear transformation on an input feature set, to produce a different feature set of lower dimensionality in a way that maximizes the proportion of the total variance that is accounted for. LDA also applies a linear transformation on the input feature set, but here the transformation is chosen to maximize a measure of class separability, and hence to improve discrimination. In order to determine the transformation, this procedure requires each input feature vector to have first been associated with a single class. PCA and LDA are both general data-reduction techniques that can usefully be applied to reduce the dimensionality of any diverse feature set, including for example static spectral or cepstral features with first- and second-order time derivatives, or even the output of auditory models. Both PCA and LDA generate new feature sets that are uncorrelated, thus allowing diagonal covariance matrices to be used for HMM state emission p.d.f.s.

10.10 VARIABLE-FRAME-RATE ANALYSIS

It has been assumed so far that all the frames in an utterance are of equal importance when making a comparison with stored templates or models. However,

Ex.1019 / Page 70 of 102
AMAZON.COM, INC.

a slight difference of vowel quality, for example, may not affect the identity of a word, whereas formant transitions at vowel–consonant boundaries may be crucial in identifying the consonant. Because, for many consonants, such transitions are very rapid, they do not occupy many frames. Although the addition of time-derivative features increases the importance of matching the transition characteristics, still rapid transitions may make only a small contribution to the cumulative distance or probability, even when they are matched very badly. The vowels and steady-state parts of long consonants can, in contrast, make a large contribution overall even when they match fairly well on each frame.

To overcome this difficulty it is necessary to give more weight to parts of the signal that are changing rapidly, and less weight to long steady regions. One way that is sometimes used to achieve this effect is to perform the original acoustic analysis at a fairly high frame rate (e.g. 100–200 frames/s), but then to discard a variable proportion of the frames depending on the distance between consecutive pairs of frames. Thus all frames are retained in rapid transitions, but perhaps only one in five is kept in very steady long vowels. This **variable-frame-rate** analysis method is similar to the scheme described in Section 4.3.5 for efficient speech coding. In the case of speech analysis for ASR, not only is there a computational saving, but also the overall match of an input utterance to stored templates or models shows much greater relative sensitivity to mismatch in transition regions.

CHAPTER 10 SUMMARY

- When deriving features for speech recognition, input speech is often first pre-emphasized by 6 dB/octave, so that the signal for subsequent analysis has a roughly flat spectral trend. Speech is analysed into a sequence of frames: most usually a 20–25 ms tapered window is applied at 10 ms intervals.
- One popular method of representing the speech spectrum is to use a filter bank with triangular filters whose width and spacing follow a mel scale. To obtain features for ASR, the output of such a filter bank is often subjected to a cosine transform (so deriving mel-frequency cepstral coefficients: MFCCs). An alternative is to derive cepstral coefficients from perceptual linear prediction.
- The cosine transform causes the features to become largely decorrelated so that diagonal covariance matrices can be used in the HMMs. In addition, information of phonetic significance is concentrated in the lower-order terms, so a more efficient representation can be obtained with fewer features.
- ASR performance is often greatly improved by adding 'delta' (first-order time-derivative) features, which are usually computed for each frame by applying linear regression over a window of five frames centred on the current frame.

CHAPTER 10 EXERCISES

E10.1 Why is cepstral analysis a useful tool in speech processing?

E10.2 Explain the stages that are typically used to analyse a speech signal into MFCCs and their first- and second-order time derivatives.

E10.3 How are properties of auditory perception simulated in front-ends for ASR?

Ex.1019 / Page 71 of 102
AMAZON.COM, INC.

CHAPTER 12

Automatic Speech Recognition for Large Vocabularies

12.1 INTRODUCTION

The previous four chapters have concentrated on introducing underlying theory and algorithms for ASR, together with some of the techniques for using the algorithms successfully in real situations. The discussion so far has deliberately concentrated either specifically on distinguishing between a small number of different words or on more general methods irrespective of the particular recognition task. In this chapter, we consider issues relevant to systems for recognizing continuously spoken utterances using large vocabularies, which may be anywhere from a few thousand up to around 100,000 different words.

12.2 HISTORICAL PERSPECTIVE

One of the earliest major efforts aimed at large-vocabulary ASR was initiated during 1971 in the United States by the Advanced Research Projects Agency (ARPA), with funding for a five-year programme of research and development. The overall objective was to make significant progress in the field of speech understanding by developing several alternative systems. The specific goal was to achieve a level of performance that was expressed in terms of semantic errors (less than 10%) on a continuous speech recognition task with a total vocabulary size of at least 1,000 words but using constrained-language input.

Although isolated-word recognition using pattern-matching techniques had achieved some initial success by the time of this ARPA programme, it was not generally obvious then how to extend the approach to accommodate the contextual effects that were known to occur in continuous speech. Therefore most systems adopted what at that time was the more traditional approach, using two separate stages. The first stage began by detecting **phonetic features** (e.g. formant frequencies, energy in different frequency bands, etc.) that were known to be important for distinguishing different speech sounds. Rules were used to convert from the measured features to a hypothesized phonetic transcription, which usually included some alternatives. The second stage then converted this transcription to a recognized word sequence. Inevitably there would be errors in the initial phonetic transcription, but the hope was that these errors would be corrected by the higher-level post-processing. However, in practice the first stage was so error-prone that information was lost which could not be recovered later. As a consequence, all the systems using this **knowledge-based** approach gave disappointing performance. In fact, the only system to achieve the required level of performance used a completely different method, based on a systematic search of a large network of

Ex.1019 / Page 72 of 102
AMAZON.COM, INC.

states with strong syntactic constraints, and it was one of the early large-vocabulary speech recognition systems using HMMs. The system was developed (somewhat separately from the main ARPA projects) at CMU by Lowerre (1976) as a Ph.D. project, extending the earlier pioneering work on HMMs by Baker (1975).

The results of the 1970s ARPA programme, while disappointing in terms of achievements for the money invested, provide a convincing demonstration of the benefits of **data-driven** statistical pattern matching over knowledge-based methods. In particular, the principle of delayed decision making is crucial, as it allows the overall best solution to be found incorporating all constraints, including those on construction of individual words and on allowed word sequences. This principle is fundamental to the design of all modern large-vocabulary speech recognizers.

Concurrent with the ARPA projects, research was in progress at IBM on the use of statistical methods for ASR. Early work was published by Jelinek (1976), independently of the work being carried out at CMU during the same period by Baker (1975). Work at IBM continued with an emphasis on applying HMMs to large-vocabulary speech recognition, and in the early 1980s the group focused on developing a system for dictation of office correspondence. The resulting system, "Tangora", as described by Jelinek (1985), was a speaker-dependent, isolated-word, near-real-time recognizer with a 5,000-word vocabulary. Although this system required users to leave pauses between words, it established the principles underlying the use of HMMs for a large-vocabulary task. Since the mid-1980s, further developments in many laboratories have led to significant further progress, and systems are now able to recognize fluent, naturally spoken continuous speech with very large vocabularies. There are a variety of systems for **large-vocabulary continuous speech recognition** (**LVCSR**) in existence, both as commercial products and as research systems in laboratories. At present, the successful systems are all based on HMMs, usually incorporating many of the refinements described in Chapter 11, but also with components that are specific to demands imposed by the need to cope with large vocabularies.

12.3 SPEECH TRANSCRIPTION AND SPEECH UNDERSTANDING

Large-vocabulary speech recognition tasks fall into two quite distinct categories:

1. *Speech transcription*: The user wishes to know exactly what the speaker said, in the form that it would be transcribed by an audio typist to produce orthographic text. Such a system may be used for dictation, and for tasks such as producing transcripts of broadcast news programmes.
2. *Speech understanding*: The semantic content of the message is required, and any recognition errors do not matter provided that the meaning is not changed. In fact often the real requirement is for the system to perform the correct action, irrespective of what words are recognized. Speech understanding systems may involve an interactive dialogue between a person and a machine to retrieve information from some computerized database. Other uses include automatic information extraction, for example to summarize spoken reports or broadcasts.

The interactive nature of many speech-understanding tasks, together with the fact that the subject area is often restricted, means that the relevant vocabulary at

Ex.1019 / Page 73 of 102
AMAZON.COM, INC.

any one point can be much smaller than the total vocabulary that is needed for more general transcription tasks. However, in order to interpret meaning of utterances, more detailed syntactic and semantic analyses are necessary than are required when just transcribing the words that were spoken. The principles of large-vocabulary recognition using HMMs apply both to transcription and to understanding, but the way in which the recognizer output is used is rather different. The first, main part of this chapter concentrates on transcription, while the latter part of the chapter briefly describes the use of large-vocabulary ASR in speech understanding systems.

12.4 SPEECH TRANSCRIPTION

The input speech waveform (typically sampled at 16 kHz) is first analysed into a sequence of acoustic feature vectors such as MFCCs (see Chapter 10). A popular choice is the first 12 cepstral coefficients and an overall energy feature together with first and second time derivatives of these features, giving a 39-element vector.

Once the input speech has been analysed into a sequence of feature vectors, the recognition task is to find the most probable word sequence $\hat{W}$ given the observed vector sequence $\boldsymbol{Y}$. Revisiting Bayes' theorem (see Section 9.2), but applying it to the task of finding a word *sequence*, the most probable sequence can be derived from the probability $P(W|\boldsymbol{Y})$ of any one sequence W as follows:

$$\hat{W} = \arg\max_{W} P(W \mid \boldsymbol{Y}) = \arg\max_{W} \frac{P(\boldsymbol{Y} \mid W)P(W)}{P(\boldsymbol{Y})} = \arg\max_{W} P(\boldsymbol{Y} \mid W)P(W) \,. \quad (12.1)$$

Equation (12.1) states that the most likely word sequence is the one which maximizes the product of $P(\boldsymbol{Y}|W)$ and $P(W)$. The first term denotes the probability of observing vector sequence $\boldsymbol{Y}$ given the word sequence W, and is determined by an **acoustic model**. The second term represents the probability of observing word sequence W independently from the acoustic signal, and is determined by a **language model**. Chapter 9 focused on the task of calculating acoustic-model probabilities, which is fundamental to any speech recognition system based on statistical models. However, for all but the most simple of applications, the language-model probability is also a major factor in obtaining good performance: restrictions imposed by the language model can greatly reduce the number of different alternatives to be distinguished by the acoustic model. As with the acoustic model, the language model for LVCSR is usually a statistical model that is automatically trained on data. In the case of the language model, these data usually take the form of *text* material chosen to be representative of the recognition task.

Assuming that models have been trained, Figure 12.1 illustrates a framework for classifying an unknown utterance by computing $P(\boldsymbol{Y}|W)P(W)$. The language model postulates a word sequence ("ten pots" in this example[1]) and determines its probability $P(W)$. In order to calculate the acoustic-model probability $P(\boldsymbol{Y}|W)$, a

[1] The phrase "ten pots" will be used in this chapter to illustrate a variety of different points. This phrase was chosen to provide a simple example for which the phonetic and orthographic transcriptions are very similar. For convenience of notation, we will represent the vowel in "pots" with its orthographic transcription /o/ in place of the correct phonetic notation for southern British English /ɒ/.

Ex.1019 / Page 74 of 102
AMAZON.COM, INC.

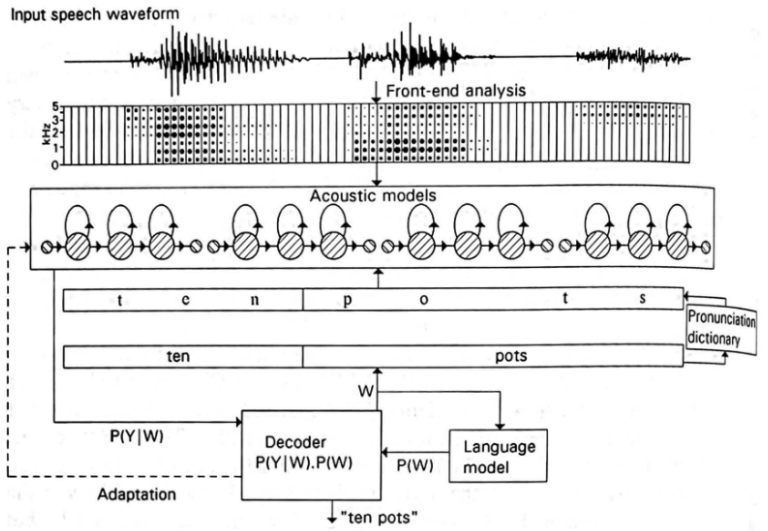


Figure 12.1 Framework for decoding a speech signal by computing the probability of a word sequence in terms of language-model and acoustic-model probabilities, shown for recognition of the phrase "ten pots". A simple filter-bank analysis is shown here for clarity of illustration, although in practice other features such as MFCCs would be used. Due to space limitations, only four of the seven models needed to represent the phone sequence in "ten pots" are shown.

composite model for the word sequence is generated. Rather than having a separate HMM for each word, the component models represent phone-size units and a pronunciation dictionary is used to specify the sequence of models for each word in the vocabulary. For any word sequence, the dictionary is used to look up the required sequence of phone models for each word, and these phone models are concatenated together to form the model for the word sequence. The probability of that model generating the observed acoustic sequence is calculated, and this probability is multiplied together with the language-model probability. In principle, this process can be repeated for all possible word sequences allowed by the language model, with the most likely sequence being selected as the recognizer output. In practice, decoding for LVCSR requires a very efficient search strategy for evaluating the vast number of different possibilities, as will be explained later.

12.5 CHALLENGES POSED BY LARGE VOCABULARIES

Issues for the design of large-vocabulary recognition systems include the following:

1. In continuous fluent speech, there are many instances when words cannot be distinguished based on acoustic information alone and it is necessary to rely on a language model for discrimination. Difficulties in making acoustic distinctions arise for two main reasons. Firstly, due to co-articulation between adjacent words, word boundaries are not usually apparent in the acoustic signal. In some cases, two utterances may be linguistically different but acoustically very similar

Ex.1019 / Page 75 of 102
AMAZON.COM, INC.

or even identical (as in the "grey day" versus "grade A" example given in Chapter 1). Secondly, the pronunciation of many words, particularly function words, can be so reduced that there is very little acoustic information at all.

2. The memory and computation requirements can become excessive. In recent years, advances in computer technology have greatly reduced the impact of this limitation. However, memory and computation are still influential, especially in determining the choice of search mechanism for use in decoding.
3. As the vocabulary size increases, it becomes increasingly difficult to provide enough representative examples of all the words, both as text to train the language model and as spoken examples to train the acoustic model.

Many of the design features of modern LVCSR systems are determined by the need to deal with these issues. The design of the acoustic model, the language model and the decoding operation are all crucial factors for the success of an LVCSR system. The following three sections describe each of these three components in turn.

12.6 ACOUSTIC MODELLING

Although some early systems used HMMs with discrete distributions for their emission p.d.f.s (e.g. Lee (1989)), current systems generally use fully continuous distributions or tied-mixture distributions, usually with diagonal covariance matrices. These latter types will be the focus of the explanation given here, which is based mainly on descriptions of the research system developed at Cambridge University (e.g. Young (1996)). This system is one of the most successful systems to date, but there are many other systems that have fairly similar structure and give broadly comparable performance, although they differ in various details.

The need to make the best use of any available acoustic training data has important consequences for the design of the acoustic-model component. With a large vocabulary, it is impractical to expect any one person to provide enough examples to train models for all the words from scratch, even if the system is intended for speaker-dependent operation. Therefore, a speaker-independent model set is used, at least to provide a starting point. Speaker-adaptation techniques are often used to improve performance for any one individual. Unsupervised adaptation may be performed using the recognizer output, as shown by the dotted data path in Figure 12.1. In addition, for a system to be used by one known person, that person can be required to speak some specific utterances, which can be used for supervised model adaptation before the person uses the system to perform any real task.

Even with several speakers to provide the data, it is not practical to train a separate model for each word in a large-vocabulary system. Even if it were practical, this approach would not make the best use of the data, as it does not take account of the fact that different words can share sub-components. Therefore large-vocabulary systems are based on **sub-word** models. The usual method, as shown in Figure 12.1, is to use models of phone-size units, with the sequence of phones for each word being specified in a pronunciation dictionary. Thus, the requirement for the training is to provide sufficient examples of all the phone-size units, and all the words in the vocabulary will not necessarily have occurred in the training data. In fact, provided suitable models are available, words can be added to the vocabulary at any time simply by extending the pronunciation dictionary.

Ex.1019 / Page 76 of 102
AMAZON.COM, INC.

12.6.1 Context-dependent phone modelling

As approximately 44 phonemes are needed to represent all English words, this number of models would be the minimum needed to build word models for English. However, the effects of co-articulation are such that the acoustic realization of any one phoneme can vary greatly with acoustic context. Therefore **context-dependent** HMMs are generally used, with different models for different phonetic contexts. Additional variation tends to arise because many speakers will, either consistently or occasionally, use word pronunciations that are different from those given in the dictionary. Although alternative pronunciations can be included in the dictionary, it is difficult to include every possible pronunciation and any that are not covered will need somehow to be accommodated in the chosen set of context-dependent HMMs.

The simplest and most popular approach is to use **triphones**, whereby every phone has a distinct HMM for every unique pair of left and right neighbours. For example, consider the word "ten". When spoken in isolation, this word could be represented by the sequence `sil t e n sil`, with the `sil` model being used for silence at the start and end. Using triphones, with the notation ${}_{x}y_{z}$ to denote phone y preceded by phone x and followed by phone z, the word would be modelled as

$$\texttt{sil}\ {}_{\texttt{sil}}\texttt{t}_{\texttt{e}}\ {}_{\texttt{t}}\texttt{e}_{\texttt{n}}\ {}_{\texttt{e}}\texttt{n}_{\texttt{sil}}\ \texttt{sil}.$$

Now consider the phrase "ten pots", for which the triphone sequence would be

$$\texttt{sil}\ {}_{\texttt{sil}}\texttt{t}_{\texttt{e}}\ {}_{\texttt{t}}\texttt{e}_{\texttt{n}}\ {}_{\texttt{e}}\texttt{n}_{\texttt{p}}\ {}_{\texttt{n}}\texttt{p}_{\texttt{o}}\ {}_{\texttt{p}}\texttt{o}_{\texttt{t}}\ {}_{\texttt{o}}\texttt{t}_{\texttt{s}}\ {}_{\texttt{t}}\texttt{s}_{\texttt{sil}}\ \texttt{sil}.$$

The two instances of the phone [t] are represented by different models because their contexts are different. Note that the triphone contexts span word boundaries, so that the first and last triphones used to represent a word depend on the preceding and following words respectively. For example, if the phrase were "ten dogs", the last triphone used to model "ten" would be ${}_{e}n_{d}$ rather than ${}_{e}n_{p}$. This use of **cross-word triphones** enables co-articulation effects across word boundaries to be accommodated, but creates complications for the decoding process as the sequence of HMMs used to represent any one word will depend on the following word.

The decoding task can be greatly simplified by using only **word-internal triphones**, whereby 'word boundary' acts as a context and so the sequence of HMMs is fixed for each word. Thus, in the above example the triphones ${}_{e}n_{p}$ and ${}_{n}p_{o}$ would be replaced by ${}_{e}n_{-}$ and ${}_{-}p_{o}$ respectively, with – being used to represent a word boundary. Early triphone systems were restricted to word-internal triphones, but the inability to model contextual effects across word boundaries is a serious disadvantage and current systems generally include cross-word context-dependent models. The consequences for decoding are explained in Section 12.8.

12.6.2 Training issues for context-dependent models

For a language with 44 different phones, the number of possible triphones is 44^3 = 85,184. In fact, phonotactic constraints are such that not all of these triphones can occur. However, an LVCSR system which includes cross-word triphones will still typically need around 60,000 triphones. This large number of possible triphones poses problems for training the models:

Ex.1019 / Page 77 of 102
AMAZON.COM, INC.

- The total number of parameters needed for the models is very large: it is usual to use three-state models with somewhere in the region of 10 mixture components to represent the output distribution for each state. This number of mixture components tends to be needed to represent the wide range of speakers (including a range of different accent types) who must be accommodated within a single model. Assuming that diagonal covariance matrices are used with 39-element acoustic feature vectors and 10 mixture components, each state would require around 790 parameters (39×10 means, 39×10 variances, and 10 mixture weights). Thus 60,000 three-state models would have a total of over 142 million parameters. Any feasible quantity of training data would not be large enough to train this number of parameters adequately.
- In any given set of training data, many triphones will inevitably not occur at all, especially if cross-word triphones are allowed (as it is very difficult to include all the phone combinations that might occur in all possible sequences of words). Thus some effective method is required for generating models for these **unseen triphones** that do not occur in the training data.

Similar issues have already been mentioned as difficulties with using whole-word models for large vocabularies. However, when using smaller model units that are meaningful in phonetic terms, it is easier to see how the problems can be reduced. The challenge is to balance the need for detail and specificity in the models against a requirement to have enough training data to obtain robust parameter estimates. To tackle the problem various different **smoothing** techniques have been used:

1. *Backing off*: When there are insufficient data to train a context-specific model, it is possible to **back off** and use a less-specific model for which more data are available. For example, one approach is to replace a triphone by the relevant **biphone**[2], representing the phone dependent on only one adjacent context, which may be either to the left or to the right. Given a choice between the left or the right biphone context, it is generally better to choose the right context as articulation tends to be anticipatory, such that following context has a greater influence than preceding context. If there are insufficient examples to train a biphone, it is possible to resort to the context-independent phone model, or **monophone**. The backing-off mechanism ensures that every model in the final system is adequately trained, but can result in only a relatively small number of full triphone models, so that several contexts are not modelled very accurately.
2. *Interpolation*: A greater degree of context dependency can be retained by **interpolating** the parameters of a context-specific (triphone) model with those of a less-specific model (such as a biphone or monophone), to give model parameters which represent some compromise between the two sets. Thus some of the context dependency of the original triphone models is preserved, while increasing their robustness by taking into account additional (less specific) data.
3. *Parameter sharing*: An alternative is to take all the triphones representing any one phone, apply some form of **clustering** procedure to group the individual

[2] Note that the term **biphone** is used to refer to a phone that is dependent upon a single context (either left or right), and that this unit is different from the **diphone** unit discussed in Chapter 5, which represents the region from the middle of one phone to the middle of the next.

Ex.1019 / Page 78 of 102
AMAZON.COM, INC.

models (or parts of models) into clusters with similar characteristics, and share the parameters between them. This sharing of parameters, often referred to as **parameter tying**, allows the data associated with similar states to be pooled to improve the robustness of the parameter estimates. This approach can retain a higher degree of context specificity than is possible with the first two methods.

Although both backing off and parameter interpolation have been used with some success, the greater power of more general parameter sharing to obtain a better compromise between accuracy and robustness is such that this method is now widely used in LVCSR. The method is described in more detail below.

12.6.3 Parameter tying

The technique of tying provides a general mechanism for sharing parameters between models or parts of models. One example is provided by the tied-mixture distributions introduced in Chapter 9, where the means and variances of each mixture component are tied across all model states. Smoothing the parameters of context-dependent models represents another application of tying. Here tying is usually applied to all model parameters for a subset of the triphones representing a phone. The aim is to tie together those models or states for which any differences in the acoustic realizations are not significant for phonetic discrimination.

Initial developments in parameter sharing between context-dependent models concentrated on clustering together triphone *models*, to give **generalized triphones**. However, this approach assumes that the degree of similarity between any two models is the same for all the states in the models. In fact, the different effects of left and right context are such that this assumption is rarely justified. For example, consider three triphones of /e/: $_{t}e_{n}$, $_{t}e_{p}$ and $_{k}e_{n}$. The first state of the $_{t}e_{n}$ and $_{t}e_{p}$ triphones can be expected to be very similar, while the last state of the $_{t}e_{n}$ and $_{k}e_{n}$ triphones will be similar. Thus tying at the state level offers much more flexibility to make the most effective use of the available data for training a set of models. We will now consider two important issues associated with the use of state tying: firstly the general procedure used to train the tied-state multiple-component mixture models (assuming that it is known which states to tie together), and secondly the choice of clustering method used to decide on the state groupings. The discussion focuses on state tying, but the principles apply in the same way when the tying is applied to complete models.

12.6.4 Training procedure

Careful design of the training procedure is essential to maximize the robustness and accuracy of the final set of tied-state context-dependent HMMs. When training sub-word models, it is not usual for the individual speech segments to have been identified and labelled in the available training data. In fact it is most likely that the data will have been transcribed as a sequence of words but not segmented at all. Rather than attempting to segment these data, they can be used directly for parameter estimation by adopting the **embedded training** approach described in Section 9.11. When using sub-word models, it is necessary first to construct a

Ex.1019 / Page 79 of 102
AMAZON.COM, INC.

model for each word from the sub-word units, before then constructing a model for the whole utterance from the individual words. Similarly to the procedure outlined in Section 12.4 for recognition, the pronunciation dictionary is used to look up the phone sequence required to represent each training utterance. A composite HMM is constructed by concatenating the appropriate sub-word models, and the relevant statistics for re-estimation are accumulated over all occurrences of each model.

Phone sequence constraints across different utterances are such that the embedded training method should generally be effective in associating appropriate speech frames with each model state, provided that in the early stages of training each model is used in a sufficient range of different contexts. For this situation it is even adequate to use the very simple 'flat' initialization of all the model parameters to identical values (see Section 9.9). It is usual to start with single-Gaussian distributions and train simple monophone models. Because there are very many examples of each one, these monophones can be trained very robustly, and provide a good basis for initializing the more specific context-dependent models. A typical procedure for training context-dependent models is illustrated in Figure 12.2, and summarized below:

1. A set of monophone HMMs, using single-Gaussian output distributions with diagonal covariance matrices, is created and trained.
2. All the training utterances are transcribed in terms of the complete set of possible triphones. For each triphone, an initial model is created by cloning the appropriate monophone. The transition probability matrix is typically not cloned, but is tied across all triphones of a phone. The triphone model parameters are re-estimated and the state occupancies, which represent the expected number of observations used to estimate the parameters of each triphone (see Section 9.5.2), are retained for later use.
3. For the set of triphones representing each phone, similar states are clustered together to create tied states (see Sections 12.6.5 and 12.6.6 for more explanation). As part of the state tying process, it is important to check that there are sufficient data associated with each tied state. This situation can be achieved by only allowing clusters for which the total state occupancy (i.e. the estimated 'count' of number of frames for which the state is occupied) exceeds a suitable threshold value (typically around 100). The parameters of the tied-state single-Gaussian models can then be re-estimated. The use of tying does not alter the form of the re-estimation formulae and can be made transparent to the re-estimation process if the data structures used to store the information are set up appropriately. Storage can be set up for accumulating the numerator and denominator for re-estimating each parameter of each tied state, with individual states simply pointing to the relevant storage.
4. Finally, multiple-component mixture models are trained using the iterative mixture splitting procedure explained in Section 9.10.4.

Delaying the introduction of the multiple-component Gaussians until the final stage has a number of advantages:

- The difficulties associated with training untied triphone mixture distributions are avoided, as multiple mixture components are only introduced once the model inventory has been set up to ensure adequate training data for each state.
- The state tying procedure is simplified because, by using single-Gaussian

Ex.1019 / Page 80 of 102
AMAZON.COM, INC.

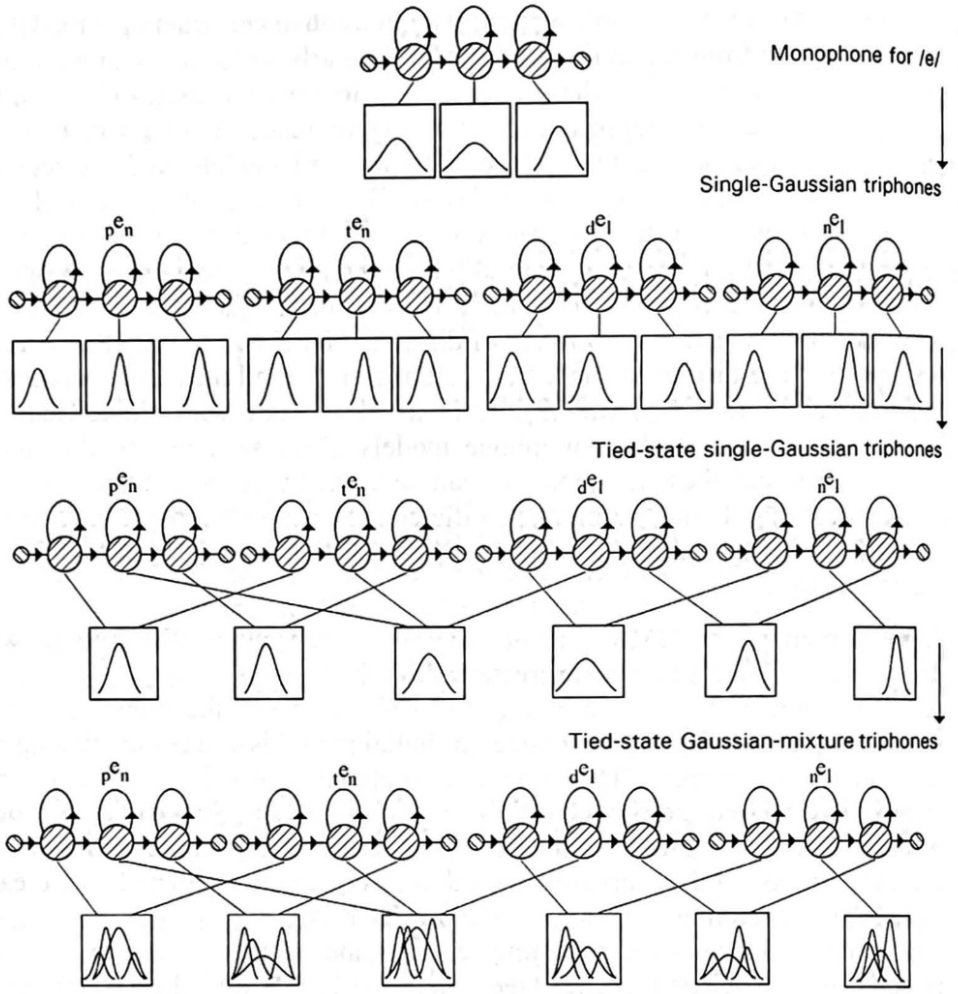


Figure 12.2 Sequence of stages for training tied-state Gaussian-mixture triphones, illustrated for a group of triphones representing the /e/ phoneme.

distributions, it is much easier to compute a similarity measure and to calculate the parameters of the combined states (see Section 12.6.6).

- By not introducing the mixture distributions at the monophone stage, the process avoids potential complications that could arise if the mixture components were used to accommodate contextual variation which would at a later stage be covered by the context-dependent models. It is generally better to accommodate contextual influences explicitly so that the predictable nature of these effects can be exploited as far as possible. The multiple mixture components are then needed mainly to allow for the fact that any one model represents data from a wide variety of different speakers.

In addition to the benefits in terms of robustness, computation and storage, state tying has the potential to lead to models with better discrimination. The potential advantages of sub-word over whole-word models in terms of discrimination power have already been mentioned. These arguments extend to the use of state tying. If the differences between the acoustic realizations associated with two different model states are simply a consequence of random variation, it is

Ex.1019 / Page 81 of 102
AMAZON.COM, INC.

detrimental for these differences to be included in the models. By combining them into a single model state, discrimination will be more focused on those regions of words containing the most useful acoustic cues. This benefit is dependent upon finding an appropriate method for determining which states to tie together.

12.6.5 Methods for clustering model parameters

Clustering methods for grouping together similar states can be divided into two general types. These methods can be used to cluster triphone states as follows:

1. *Bottom-up clustering*: Starting with a separate model for each individual triphone, similar states are merged together to form a single new model state. The merging process is continued until some threshold is reached which ensures that there are adequate data to train each new clustered state. This data-driven approach is often referred to as **agglomerative clustering**. The method should ensure that there are sufficient data to train every state in the final set, while keeping the models as context-specific as possible given the available training data. However, for any triphones that do not occur at all in the training data, it is still necessary to back off to more general models such as biphones or monophones.
2. *Top-down clustering*: Initially all triphones for a phoneme are grouped together and a hierarchical splitting procedure is used to progressively divide up the group according to the answers to binary yes/no questions about either the left or the right immediately adjacent phonetic context. The questions are arranged as a **phonetic decision tree**, and the division process starts at the root node of the tree and continues until all the leaf nodes have been reached. All the states clustered at each leaf node are then tied together. A tree is generated for each state of each phone. An example showing the use of a decision tree to cluster the centre state of some /e/ triphones is shown in Figure 12.3. The context questions in the tree may relate to specific phones (e.g. "Is the phone to the right /l/?"), or to broad phonetic classes (e.g. "Is the phone to the left a nasal?"). Using the tree, the correct tied state to use for any given context can be found by tracing the path down the tree until a leaf node is reached (see Figure 12.3).

The main advantage of the top-down approach to clustering is that a context-dependent model will be specified for *any* triphone context, even if that context did not occur in the training data. It is thus possible to build more accurate models for unseen triphones than can be achieved with the simple backing-off strategy, assuming that the questions in the tree are such that contexts are grouped appropriately. Although the tree could be constructed by hand based on phonetic knowledge, this approach does not work very well in practice, as it does not take into account the acoustic similarity of the triphones in the data. It is, however, possible to construct trees automatically by combining the use of phonetic questions with tests of acoustic similarity and a test for sufficient data to represent any new division. This automatic construction provides generalization to unseen contexts while maintaining accuracy and robustness in the acoustic models. A popular version of this effective technique for constructing phonetic decision trees is explained in more detail in the next section.

Ex.1019 / Page 82 of 102
AMAZON.COM, INC.

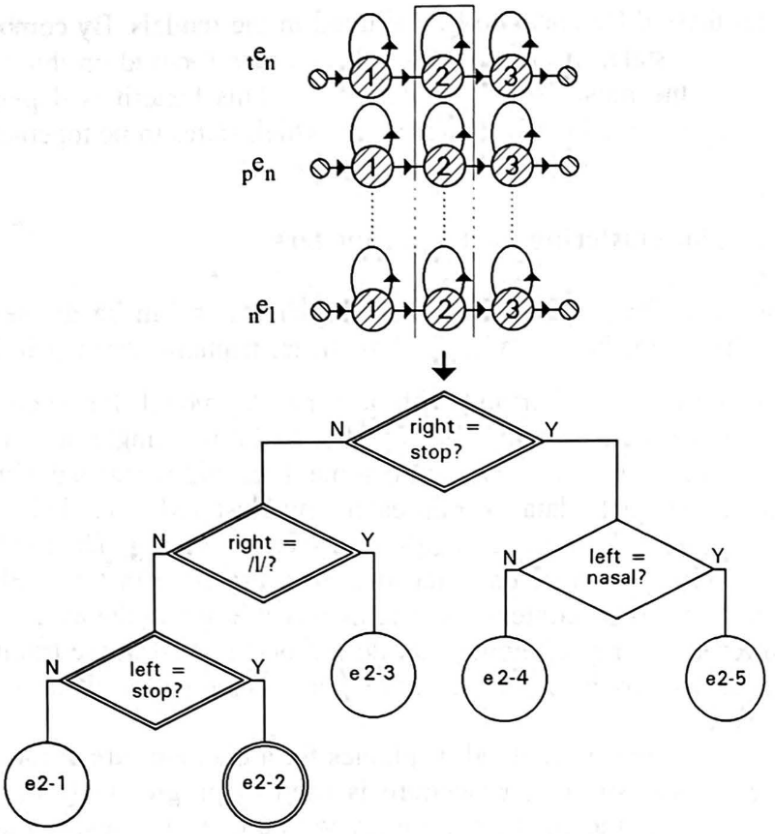


Figure 12.3 Example illustrating the use of a phonetic decision tree to cluster the centre state for a group of triphones of the /e/ phoneme. Each triphone is moved down the tree until a terminal 'leaf' node is reached (shown as circles), and a new model state is formed from the members of the cluster. The tree shown here is a very simple one intended simply as an illustration of the principles of clustering, and in any real application there will be many more questions before a terminal node is reached. The tree can then be used to find the appropriate clustered state for any given context. As an example, the route down this simple tree is shown (using double lines round the chosen decision boxes and the final leaf node) for the context of a preceding /t/ and a following /n/. In this context the clustered state e2-2 will be used.

12.6.6 Constructing phonetic decision trees

First, linguistic knowledge is used to choose a set of possible context questions that might be used to divide the data. This question set will usually include tests for each specific phone, tests for phonetic classes (e.g. stop, vowel), tests for more restricted classes (e.g. voiced stop, front vowel) and tests for more general classes (e.g. voiced consonant, continuant). Typically, there will be over 100 questions for the left context and a similar number for the right context. For each state of each phone, the aim is to build a tree where the question asked at each node is chosen to maximize the likelihood of the training data given the final set of tied states. A condition is imposed to ensure that there are sufficient data associated with each tied state (i.e. that the total occupancy of the tied state exceeds some threshold).

It would in principle be possible to build all possible tree architectures (for all states), train a set of models for each architecture, and choose the set for which the likelihood of the training data is the highest while satisfying the occupancy

Ex.1019 / Page 83 of 102
AMAZON.COM, INC.