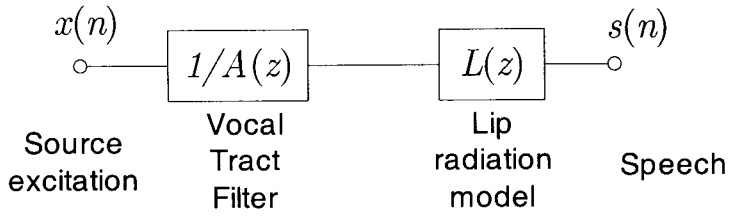
Ise 4.17

S016

Solutions

Speech Processing 2003

1. a)



b)

Recognition: We want parameters that are good at distinguishing between different speech sounds. I.e. different speech sounds should give different parameter values whereas multiple examples of any particular speech sounds should give similar values.

Coding: Quantisation errors must not affect the spectrum too much. It is best if parameters have a natural order (unlike for example the pole positions) as this means the transmission order conveys useful information and a particular parameter will have less variability. Want easy stability check. Want to interpolate between frames

(c)

We find roots at $0.49 \pm 0.8487j$ leading to a frequency of $\pi / 3$ and a gain of 29.2953 dB.

The gain is given by:

$$= -10\log_{10}\left((1 - a_1 z^{-1} - a_2 z^{-2})(1 - a_1 z - a_2 z^2)\right)$$

$$= -10\log_{10}\left(1 + a_1^2 + a_2^2 + a_1(a_2 - 1)(z + z^{-1}) - a_2(z^2 + z^{-2})\right)$$

for the specific case of fs/6 given, $z+z^{-1} = 1$ and $z^2+z^{-2} = -1$, hence the gain is

$$= -10\log_{10}\left(1 + a_1^2 + a_2^2 + a_1(a_2 - 1) + a_2\right)$$

$$= -10\log_{10}\left((1 - a_1)(1 + a_2) + (a_1 + a_2)^2\right)$$

$$= 29.3 \text{ dB}$$

For a 1% increase in a_2, the gain becomes 31.55 dB.

(d)

Use the conversion formulae

$$r_1 = \frac{-a_1}{1 - a_2}$$

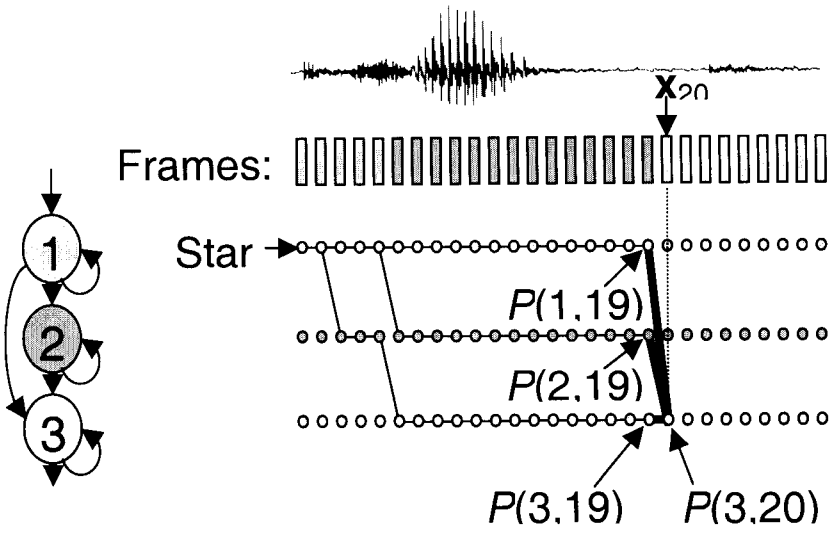
$$r_2 = -a_2$$

so that r1 = -0.4999 and r2 = 0.9604.

From the expression $-a1 = r1 + r1r2$ it can immediately be seen that a 1% change in r1 will have exactly the same effect as a 1% change in a1. A 1% change in r2 gives a new gain value of 31.7 dB, ie. a worse robustness than the prediction coefficients. The hypothesis is therefore false for this filter.

Speech coding always introduces errors in the parameters. Robustness to such errors is essential for low bit-rate schemes. The features normal showing the most robustness are LSPs.

2. (a) (i)



In this example, any alignment going through (3,20) must go through either (1,19), (2,19) or (3,19). Hence:

$$P(3,20) = \left(\sum_{s=1}^{3} P(s,19) \times a_{s3} \right) \times d_3(\mathbf{x}_{20})$$

This can be generalized to say that

$$P(k,t) = \left(\sum_{s=1}^{S} P(s,t-1) \times a_{sk} \right) \times d_k(\mathbf{x}_t) .$$

By a similar argument

$$Q(t,s) = \sum_{i=1}^{S} a_{si} \times d_i(\mathbf{x}_{t+1}) \times Q(t+1,i)$$

The initial conditions for the recursions are:

$P(1,s) = d_1(\mathbf{x}_1)$ for $s = 1$ else $= 0$

$Q(T,s) = a_{S=}$ for $s = S$ else $= 0$

(ii)

$P(1,1) = d_1(\mathbf{x}_1)$; $P(s,1) = 0$ for $s \neq 1$

for t=2:T

for k=1:S

$$P(k,t) = \left(\sum_{s=1}^{S} P(s,t-1) \times a_{sk} \right) \times d_k(\mathbf{x}_t)$$

end

end

(b)

We calculate the Q values recursively as follows:

$$Q(5,3) = 0.8$$

$$Q(5,2) = Q(5,1) = 0$$

$$Q(4,3) = 0.2 \times 0.5 \times Q(5,3) = 0.08$$

$$Q(4,2) = 0.5 \times 0.5 \times Q(5,3) = 0.2$$

$$Q(4,1) = 0$$

And the P values:

$$P(1,1) = 0.5$$

$$P(1,2) = P(1,3) = 0$$

$$P(2,1) = P(1,1) \times 0.9 \times 0.1 = 0.045$$

$$P(2,2) = P(1,1) \times 0.1 \times 0.4 = 0.02$$

$$P(3,1) = P(2,1) \times 0.9 \times 0.8 = 0.0324$$

$$P(3,2) = P(2,1) \times 0.1 \times 0.2 + P(2,2) \times 0.5 \times 0.2 = 0.0029$$

$$P(3,3) = P(2,2) \times 0.5 \times 0.5 = 0.005$$

$$P(4,1) = P(3,1) \times 0.9 \times 0.6 = 0.017496$$

$$P(4,2) = P(3,1) \times 0.1 \times 0.3 + P(3,2) \times 0.5 \times 0.3 = 0.001407$$

$$P(4,3) = P(3,2) \times 0.5 \times 0.2 + P(3,3) \times 0.2 \times 0.2 = 0.00049$$

We need $P(4,i) \times Q(4,i)$ for $i = 1,2,3$. This is 0, 0.000281 and 0.0000392 respectively.

3.(a)

(i) Markov model: The probability of being in any state for the next frame depends only on the current state and not on any previous history. The alignment of frames-to-states is hidden.

(ii) Left to right, no skips: all transition probabilities are zero except for transitions to the next (left to right) state in the model. This implies that the model is not capable of entering a state that has previously been used, and must include at least one frame in each state.

(iii) Continuous density: the parameters describing each state are drawn from a continuous probability density function, typically Normal and characterized by mean and variance. This is in contrast to discrete models in which the parameters are taken from a discrete set, typically formed from the application of VQ to the input features.

(b)

$$\mathrm{pr}(D=n) = p(1-p)^{n-1} \Rightarrow \mathrm{E}(D) = \sum_{n=1}^{\infty} np(1-p)^{n-1} = \frac{1}{p}$$

$$\text{Since by differentiating } \sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \Rightarrow \sum_{n=0}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

(c)

When different people say the same phoneme, the feature vectors should have similar values.

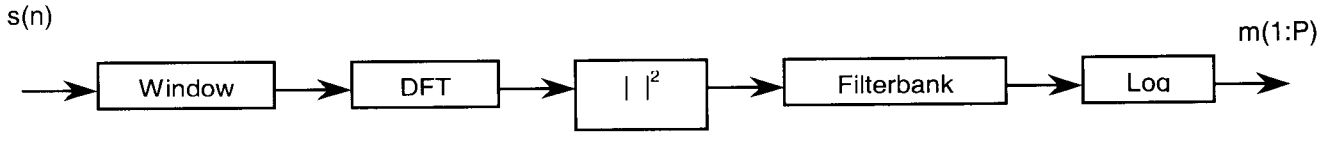
Different phonemes from the same or different speakers should give dissimilar values.

For different examples of the same phoneme, the features should be independent and uncorrelated: this allows us to multiply their probabilities.

For different examples of the same phoneme, each feature should preferably follow a probability distribution that is well described as a sum of gaussians.

The features should not be affected by the amplitude of the speech signal otherwise recognition performance would vary with your distance from the microphone.

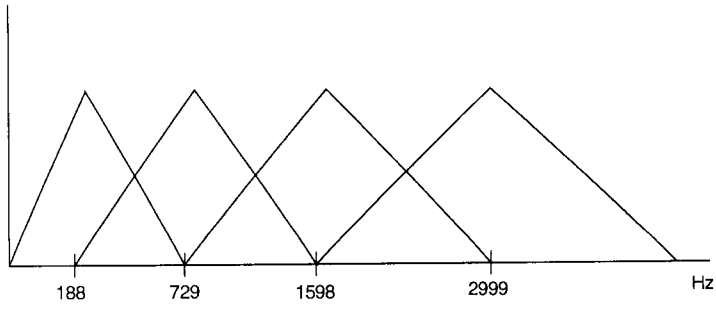
(d) Mel-cepstrum coefficients:



For the Mel filterbank to be uniformly spaced:
min freq = 0 Hz = 0 mel.
max freq = 4000 Hz = 2146 mel

For 4 uniformly spaced filters, centres as follows:

Mel	268	804	1340	1876
Hz	188	729	1598	2999



4. (a) The sequence of steps is:

→ Words	Convert symbols, abbreviations, acronyms, dates and numbers into words. For some abbreviations, the meaning changes with context: e.g. St Peter = Saint but Peter St = Street. Some acronyms are pronounced as words (e.g. NATO) whereas others are spelled out as individual letters (e.g. HIV). The pronunciation of integers changes when they are dates: "1066 people died at the battle of Hastings in 1066".
→ Phonemes	Convert each word to its constituent phonemes using a combination of a word dictionary, morph decomposition and letter-to-sound rules. There are many word-pairs with the same spelling but different pronunciation (e.g. bow=bend at the waist, bow=knot tied in a ribbon): it is extremely difficult to choose between such pairs reliably.
Parsing	It is necessary to parse the sentence to decide where phrase boundaries occur. This is essential in order to apply the right pauses and durations to the speech.
+ Pauses/Durations	Insert pauses between sentences and phrases. Puses are also required within long phrases to allow the "speaker" to take a breath. Phonemes are lengthened prior to a pause.
+ Stress and Pitch	Specify the amplitude of each phoneme and its duration taking account of the sentence structure and rhythym. Each word will generally have a primary stress position identified during word-to-phoneme conversion. The parsing of the sentence will further indicate which words within the sentence should be stressed. The pitch of a sentence normally rises to the first stressed word and then falls. Yes/No questions rise in pitch at the end unless they begin with "Wh...".
→ Waveform	Convert the phoneme string to a waveform using formant synthesis or diphone/demisyllable synthesis. In either case it is necessary to ensure smoothly changing pitch and amplitude contours. With diphone or demisyllable synthesis this can be achieved in the time-domain using the PSOLA algorithm. With formant synthesis, this is automatic.

(b) A dictionary is essential for irregular words (e.g. of, meringue)

Decomposing words into morphemes allows a much smaller dictionary to be used (because there are fewer morphemes than words) and will also cope sensibly with newly coined words, which are normally built from existing morphemes. The morpheme decomposition is able to give grammatical information about even newly formed words: this helps with prosody determination.

(c) There are five possible decompositions (P,R and S denote Prefix, Root and Suffix respectively):

Decomposition	**Elements**	**Cost**
un+in+form+ed	PPRS	204
un+inform+ed	PRS	170
un+in+formed	PPR	169
un+in+form+ed	PRRS	303
un+in+formed	PRR	268

5.
For covariance LPC we have

$$\sum_{j=1}^{p} \phi_{ij} a_j = \phi_{i0} \quad \text{where} \quad \phi_{ij} = \sum_{n \in \{F\}} s(n-i)s(n-j)$$

or in matrix for $\mathbf{\Phi a} = \mathbf{c}$ with the vector **c** defined by $c_i = \phi_i$.

We chose *{F}* to be a finite segment of speech. $\phi_{ij} = \sum_{n=0}^{N-1} s(n-i)s(n-j)$

The matrix Φ is symmetric but is no longer Toeplitz. It can be derived recursively as

$$\begin{aligned}
\phi_{ij} &= \sum_{n=-1}^{N-2} s(n-i+1)s(n-j+1) \\
&= s(-i)s(-j) - s(N-i)s(N-j) + \sum_{n=0}^{N-1} s(n-i+1)s(n-j+1) \\
&= s(-i)s(-j) - s(N-i)s(N-j) + \phi_{i-1,j-1}
\end{aligned}$$

The speech must first be divided into frames of sufficient duration (>2ms) but not so long as to violate the assumption of stationarity within a frame (eg. <30ms). Then the LPC coefficients are found for each frame by solution of the above equations. Then format frequencies can be found by factorizing the predictor polynomial. Some pole pairs will correspond to well-defined formants and some to general spectral shaping. The latter cases can be identified by their wide bandwidth (ie. they are much closer to the origin in z).

Replacing the summation by an expectation:

$$\begin{aligned}
\phi_{ij} &= E(s(n-i)s(n-j)) \\
&= E(\sin(2\pi f(n-i))\sin(2\pi f(n-j))) + \sigma^2\delta_{ij} \\
&= \tfrac{1}{2}E(\cos(2\pi f(-i+j)) - \cos(2\pi f(2n-j-i))) + \sigma^2\delta_{ij} \\
&= \tfrac{1}{2}\cos(2\pi f(j-i)) + \sigma^2\delta_{ij} \\
c_i &= \phi_{i0} = \tfrac{1}{2}\cos(2\pi f i)
\end{aligned}$$

For the case, *f* = 0.1 and $\sigma^2 = 0.05$ we have

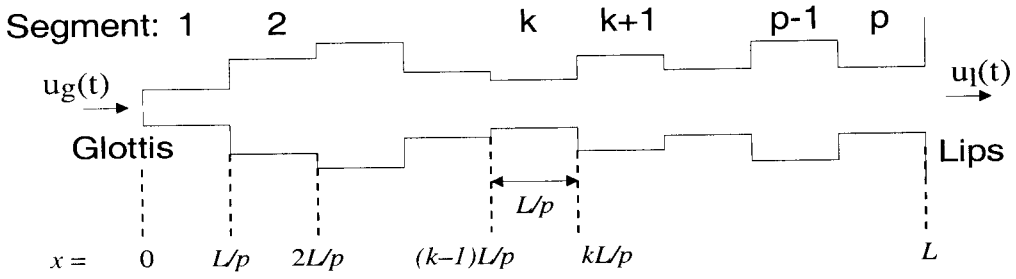
$$\begin{aligned}
\phi_{ij} &= E\left(\sin(2\pi f(n-i)\sin(2\pi f(n-j)) + \sigma^2\delta_{ij}\right) \\
&= \frac{1}{2}E\left(\cos(2\pi f(-i+j)) - \cos(2\pi f(2n-i-j) + \sigma^2\delta_{ij}\right) \\
&= \frac{1}{2}\cos(2\pi f(j-i)) + \sigma^2\delta_{ij} \\
\phi_{i0} &= \begin{bmatrix} \frac{1}{2}\cos(2\pi \times 0.1) \\ \frac{1}{2}\cos(2\pi \times 0.2) \end{bmatrix}
\end{aligned}$$

Therefore

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.55 & 0.4045 \\ 0.4045 & 0.55 \end{pmatrix}^{-1} \begin{pmatrix} 0.405 \\ 0.155 \end{pmatrix} = \frac{1}{0.139} \begin{pmatrix} 0.55 & -0.4045 \\ -0.4045 & 0.55 \end{pmatrix} \begin{pmatrix} 0.405 \\ 0.155 \end{pmatrix} = \begin{pmatrix} 1.152 \\ -0.566 \end{pmatrix}$$

The roots of $1-a_1 z^{-1}-a_2 z^{-2}$ are $z=0.576\pm 0.484\,j=0.753\angle 0.699$ giving a frequency estimate of $\frac{1}{2\pi}\cos^{-1}\left(\frac{a_1}{2\sqrt{-a_2}}\right)=0.113$, an error of 11% at 10 dB SNR.

6. We model the vocal tract as a tube that has p segments:

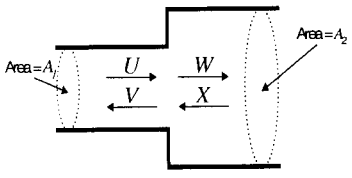


u_g and u_l are the volume flows of air at the glottis and lips respectively (measured in litres per second).

Vocal tract is of length L (typically 15-17 cm in adults).

Length of each segment is the distance sound travels in half a sample period = $0.5cT$: 1.5 cm @ 11 kHz
c = speed of sound in air
T = sample period = $1/f_{samp}$

Number of tube segments needed = $2L/cT$ ~= 0.001 f_{samp}



Flow Continuity: $(U - V) = (W - X)$; Pressure Continuity: $\frac{\rho c}{A_1}(U + V) = \frac{\rho c}{A_2}(W + X)$

Reflection coefficient: $r = \frac{A_2 - A_1}{A_2 + A_1}$

Hence:

$$\begin{pmatrix} U \\ V \end{pmatrix} = \frac{1}{2A_2}\begin{pmatrix} A_2 & 1 \\ -A_2 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ A_1 & A_1 \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix} = \frac{1}{2A_2}\begin{pmatrix} A_1 + A_2 & A_1 - A_2 \\ A_1 - A_2 & A_1 + A_2 \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

$$= \frac{1}{1+r}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

$$U = \frac{1}{1+r}[W - rX]$$

$$V = \frac{1}{1+r}[-rW + X]$$

$$\therefore W = (1+r)U + rX, \quad V = -rU + (1-r^2)\frac{X}{1+r}$$

$$\begin{bmatrix} V \\ W \end{bmatrix} = \begin{bmatrix} -r & 1-r \\ 1+r & r \end{bmatrix}\begin{bmatrix} U \\ X \end{bmatrix}$$

For the 2 tube model we have a transfer function of

$$\frac{Gz^{-1}}{1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2}}$$

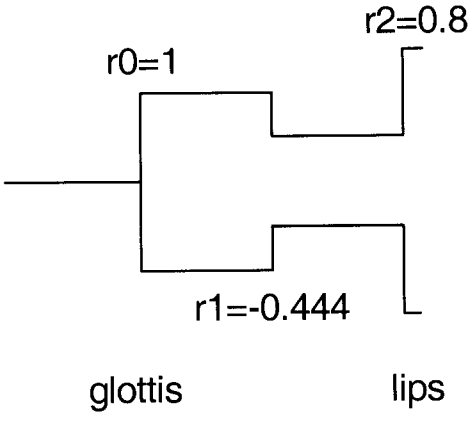
Comparing terms we obtain

$$G = 6$$
$$r_0r_1 + r_1r_2 = -0.8$$
$$r_0r_2 = 0.8$$

If the glottis is closed this implies $r_0 = 1$. Hence

$r_2 = 0.8$

$r_1 = -0.444$



Non-ideal characteristics include vibrational losses in the vocal tract walls, viscous friction and thermal conduction.