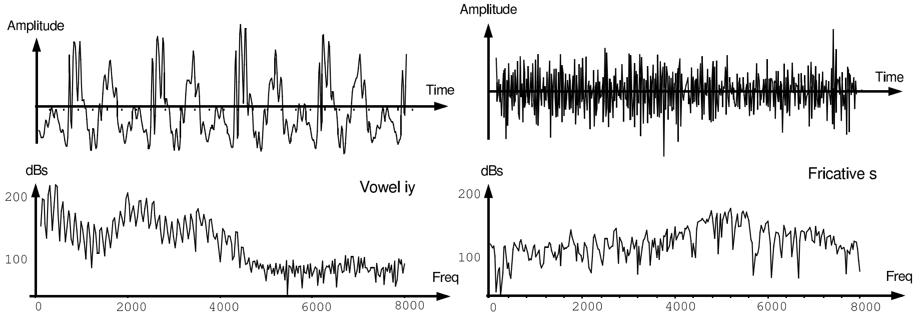
Solutions to 4F11 Speech Processing, 2003

1. *Speech Analysis*

(a)(i) The source filter models makes a clear distinction between voiced and unvoiced signals. In these cases the excitation source is either a noise or a sequence of pulses. [10%]

(a)(ii) The waveform of a voiced speech segment shows the periodic nature. This is reflected in the spectrum as a ripple with peaks at multiples of the fundamental frequency (F0). Further clear resonances are visible. In contrast the unvoiced signal is noise-like and the associated magnitude spectrum does not reveal any periodicity. Only weak and broad resonances can be observed. Comparing voiced and unvoiced signals also exhibits a substantial difference in energy.



[25%]

(b)(i) A linear predictor tries to predict the value of the signal s at time n on the basis of the past p signal values, where p is the predictor order. It does so by a linear combination with the weight factors a_i which are called the linear prediction coefficients.

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i}$$

[15%]

(b)(ii) A least squares error criterion is used: The error signal can be computed as the difference of the predictor output and the true signal value:

$$e_n = s_n - \hat{s}_n = s_n - \sum_{i=1}^{p} a_i s_{n-i}$$

Thus for obtaining the optimal filter coefficients a_i requires to take the derivative of

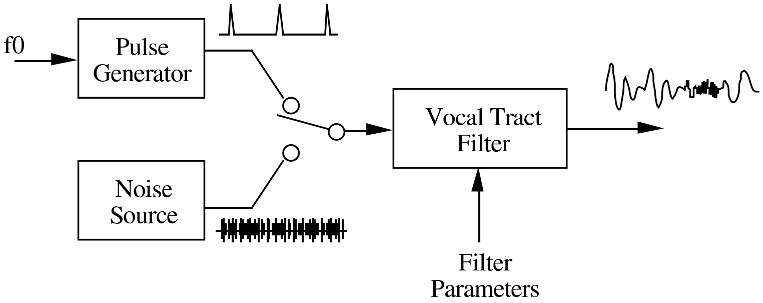
$$E = \sum_n \left(s_n - \sum_{i=1}^{p} a_i s_{n-i} \right)$$

with respect the parameters $a_i$and setting the derivative to 0. The minimisation of the error signal is equivalent to minimising the expression

$$\int_{-\pi}^{\pi} \frac{P(e^{j\omega T})}{\hat{P}(e^{j\omega T})} d\omega T$$

where $P(\hat{e^{j\omega T}})$ represents the LP power spectrum and $P(e^{j\omega T})$ is the speech power spectrum (under the constraint that the overall energy is the same). A better match is obtained for large values of the spectrum, i.e. the peaks, and less emphasis is given to the matching of valleys. [25%]

(c) The source-filter model is outlined below:



with the vocal tract filter represented by an LP filter. In order to synthesise a speech signal a selection is made whether the current signal is voiced or unvoiced. In the case of a voiced signal a sequence of impulses on the basis of the fundamental frequency is generated, in the unvoiced case a random noise signal is used. After multiplication with a gain factor the signal filtered using the LP filter. In order to reflect the rapid changes in the speech signal the parameters need to be updated every 10ms. The limitations are

- A hard V/UV decision is invalid for example in the voiced fricative case → mixed excitation.
- Neither random noise nor periodic pulses represent the true signal sources correctly.
- Nasals are not properly modelled with an LP filter
- F0 estimation or synthesis is difficult
- ...

[25%]

2. *Speech Coding*

(a) Design criteria are

- of the effective *bit-rate*, which may be variable or constant,
- the *speech quality*

- the *complexity* of the encoding and decoding algorithms
- the delay incurred due to processing
- the *generality* for use on other types of audio signals

[20%]

(b) The factors are the assessment of *intelligibility* of the speech signal and the *naturalness*. Techniques used to assess the quality of a speech coder are objective noise measurements such as a segmental Signal-to-noise ratio or subjective listening tests. For assessment of the quality of speech signals encoded with parametric techniques only subjective evaluation is of importance. Here the *Diagnostic Rhyme Test* allows to measure confusion between phones by asking whether for example `hit` or `fit` was uttered. The more widely used Mean Opinion score tests mostly for naturalness by requiring the assignment of scores from 1 to 5 to a particular speech sample. In both tests it is unknown to the listener if the presented speech sample is encoded or not. [30%]

(c)(i) The parameters are the gain G, the fundamental frequency F_0, a V/UV decision and the parameters associated with the LP filter. The standard uses 5 bits for G, 6 bits for F_0, 1 bit for V/UV and 42 bits for encoding the LP parameters (mixed encoding of log-area-ratios and reflection coefficients). [15%]

(c)(ii) The above is a total of 54 bits , transferred every 180 samples at a sample rate of 8000Hz: $\frac{8000}{180} \times 54 = 2400\text{bit/s}$ [10%]

(c)(iii) Another option to lower the bit-rate is to use quantisation of the individual filter coefficients. The filter coefficients show strong correlation between each other. In order to exploit the correlation between the individual filter parameters vector quantisation can be used. In this a codebook is trained on a large training set of LP parameter vectors using for example k-means clustering. This also requires the choice of an appropriate distance metric. Especially for computing the distance between LP parameter vectors the Itakura distance can be used. In encoding the centroid vector nearest to the LP parameter vector is chosen and the index is transmitted. The coder uses the same codebook and uses the centroid LP parameter vector decoding [25%]

3. *Frontend Extraction for SPeech Recognition*

(a) The desirable attributes are

- Reduce the raw bit rate to something manageable .
- Remove information that does not discriminate between words
- Retain all information that disriminates between words
- Transform feature vector to be suitable for the classifier being used.

[20%]

(b) The steps in producing an MFCC feature vector are:

- Pre-emphasis of the speech signal
  This provides an attenuation of lower frequency components to compensate for the tilt in the speech spectrum.

- Block processing, i.e. taking a frame of the speech signal every 10ms.
  This allows to use the assumption of quasi-stationary speech segments
- Windowing using a Hamming window of 25ms
  This introduces less distortion (side-lobes) than a rectangular window.
- DFT of the windowed signal to obtain the magnitude spectrum
- Filtering using a triangular filterbank on the basis of the Mel scale. A typical number of filters is 26.
  This provides a smoothed representation of the spectrum while taking into account that the speech in lower frequency regions is perceptually more important.
- Take the log of the filterbank coefficients
  This converts the multiplicative relationship between excitation and source into an additive one.
- Take the inverse DCT of the log filterbank coefficients. Only retain a smaller number of MFCC elements, a typical value is 12.
  This steps computes the cepstral values, further smoothing the spectrum by truncation.

[35%]

(c) The delta and delta-delta coefficients allow to incorporate dynamic information about neighbouring feature vectors in time into the current vector. Define a static feature vector $\mathbf{y}$ (for example 12 MFCCs as outlined in (b)), then the delta parameters are computed by

$$\frac{\sum_{\tau=1}^{D} \tau \left(\mathbf{y}_{t+\tau} - \mathbf{y}_{t-\tau}\right)}{2\sum_{\tau=1}^{D} \tau^2}$$

which is a linear regression using $2D$ data points. This yields 12 delta coefficients. The delta-delta or acceleration coefficients are obtained by computing the regression values on the first order differentials.

The encoding of temporal information into the feature vector is commonly used to counteract the conditional independence assumption between subsequent observation vectors, as used in HMMs.

The size of the feature vector is substantially increased (three-fold) using this technique, consequently increasing the computational complexity.

[25%]

(d) This was not detailed in lectures. A selvction from the following points will obtain full marks.

- Background noise n_k is added to the signal

$$\tilde{s}_k = s_k + n_k$$

- Most of the operations in (b) are linear filter operations in which case the principle of superposition can be used to predict the change to the coefficients. This is true for pre-emphasis, windowing, and DFT, and the filterbank. Up to this stage the two parts of the signal are still additive.

- The output of the filterbank is then passed through a log. This is non-linear so the speech and noise can no longer be considered seperatly.
- The DCT means that all filterbank outputs affect one another. Thus if the noise is band limited then only a limitred number of the filterbank valuse will be corrupted, whereas all the cepstral values will be altered.

[20%]

4. *Hidden Markov models*

(a) The two systems are

- System (i)

$$\log(p(\boldsymbol{y})) = \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\right) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{1}$$

  Number of model parameters $d + \frac{d}{2}(d+1)$. Able to model unimodal data, but with correlations in the feature vector. Calculation cost is $\mathcal{O}(d^2)$.

- System (ii)

$$\log(p(\boldsymbol{y})) = \log\left(\sum_{m=1}^{M} \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_m|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_m)'\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_m)\right)\right) \tag{2}$$

  The number of model parameters is $M(2d+1)-1$. Able to model multi-modal data, on-Gaussian data and to a limited extent correlations.

As the size of the feature vector increases, the full-covariance system increaes as $\mathcal{O}(d^2)$, for the multiple component system $\mathcal{O}(d)$. [35%]

(b)(i) Cross-word triphone systems makes the model to be used dependent on the preceeding and following phones, as well as the current. Commonly used as the articulators do not move instantly, so there is significant co-articulation. [15%]

(b)(ii) The problems are:

- how to construct models when there is no observed training data;
- how to obtain robust estimates of model parameters when there is little data.

Main attributes are:
Advantages

- No need to back-off, smooth use made of contextual information.
- Allows expert knowledge to be incorporated
- Allows any degree of context dependency to be simply incorporated.

Disadvantages

- Requires expert knowledge to specify question set
- Locally optimal decisions made
- Not all possible question combinations asked

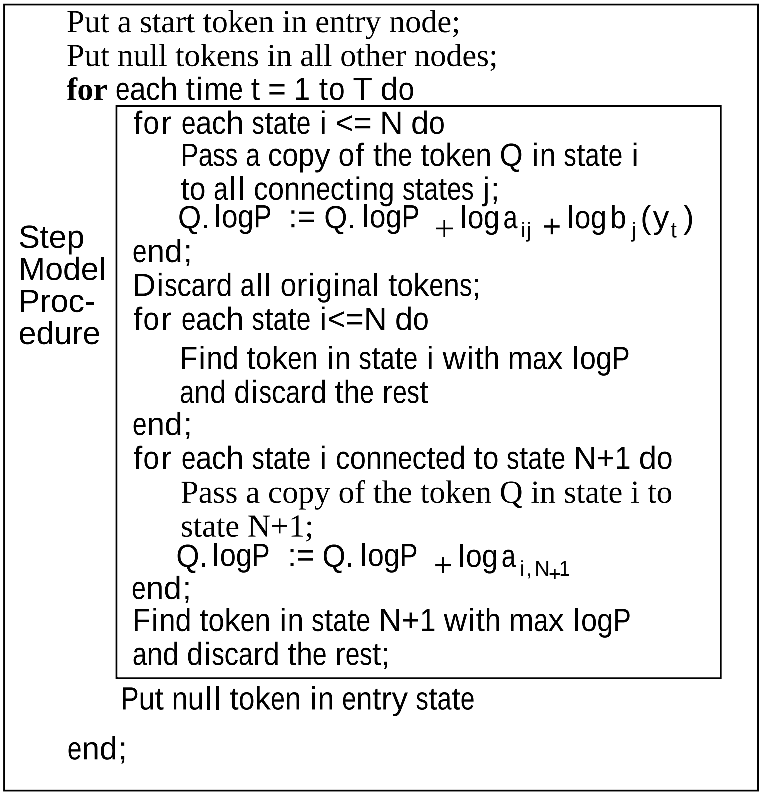
[30%]

(c) The GMM system uses multiple components to model each state. The context may be implicitly modelled using the GMM. The modelling is explicit for the triphone system. Problem is that the GMM system allows the "context" to swap every frame. It also allows non-Gaussian and bimodal distributions to be modelled. [20%]

5. *Search*

(a) Token passing can be taken straight from notes. Define

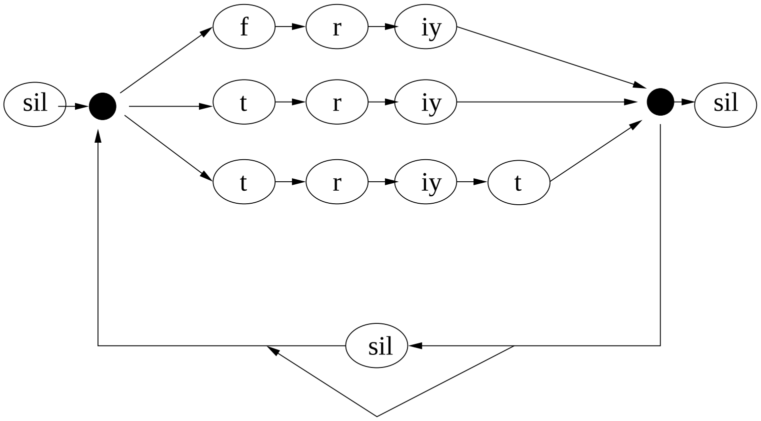
- a **start token** to be a token `Q` such that `Q.logP` is zero (i.e. pprobability = 1.0)
- a **null token** to be a token such that `Q.logP` is negative infinity (i.e. pprobability = 0.0).



```
Put a start token in entry node;
Put null tokens in all other nodes;
for each time t = 1 to T do
Step Model Procedure
    for each state i <= N do
        Pass a copy of the token Q in state i
        to all connecting states j;
        $Q.\log P := Q.\log P + \log a_{ij} + \log b_j(y_t)$
    end;
    Discard all original tokens;
    for each state i<=N do
        Find token in state i with max logP
        and discard the rest
    end;
    for each state i connected to state N+1 do
        Pass a copy of the token Q in state i to
        state N+1;
        $Q.\log P := Q.\log P + \log a_{i,N+1}$
    end;
    Find token in state N+1 with max logP
    and discard the rest;
    Put null token in entry state
end;
```

[25%]

(b) The linear lexicon is given over the page.



[20%]

(c) (i) Every HMM state is either *active* or *inactive*. Initially, all network entry states are active and all other states are inactive. The beam search algorithm then works as follows

(a) For each active state

  i. Propagate internal tokens as normal

  ii. Record max `logP` in any state, call this `gMax`

(b) De-activate all states for which `logP < gMax - B` where `B` is the beam width

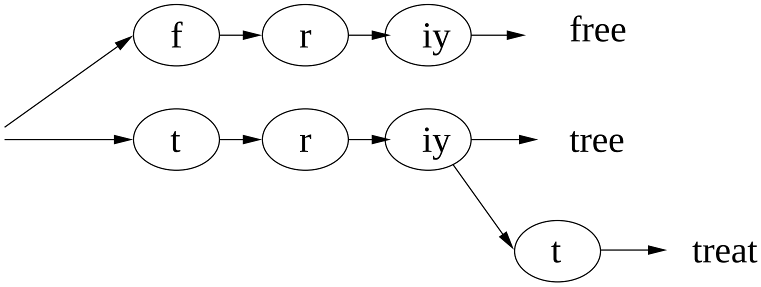
(c) Propagate external tokens only if they are within the beam i.e. `logP < gMax - B`

(d) Re-activate all states which have received a new entry token

[20%]

(c) (ii) For efficient implementation the unigram lannguage model is added at the start of each word. This allows all information source to be incorporated as soon as possible. [15%]

(d) The tree-structured lexicon is given below



This reduces the number of active paths at the start of all the words. Since most paths from incorrect words are pruned out after the first few frames, this dramatically reduces the number of path propagations required. [20%]